ACL-IJCNLP 2009 Handbook

Joint Conference of the 47^{th} Annual Meeting of the
Association for Computational Linguistics
and
the 4^{th} International Joint Conference on Natural Language Processing of the
Asian Federation of Natural Language Processing

2 – 7 August, 2009
Suntec, Singapore

Organizing Committee

General Conference Chair
Keh-Yih Su (Behavior Design Corp., Taiwan)

Program Chairs
Jian Su (Institute for Infocomm Research, Singapore)
Janyce Wiebe (University of Pittsburgh, USA)

Local Organizing Chair
Haizhou Li (Institute for Infocomm Research, Singapore)

Demo Chairs
Gary Geunbae Lee (POSTECH, Korea)
Sabine Schulte im Walde (University of Stuttgart, Germany)

Exhibits Chairs
Timothy Baldwin (University of Melbourne, Australia)
Philipp Koehn (University of Edinburgh, UK)

Mentoring Service Chairs
Hwee Tou Ng (National University of Singapore, Singapore)
Florence Reeder (Mitre, USA)

Publication Chairs
Regina Barzilay (MIT, USA)
Jing-Shin Chang (National Chi Nan University, Taiwan)

Publicity Chairs
Min-Yen Kan (National University of Singapore, Singapore)
Andy Way (Dublin City University, Ireland)

Sponsorship Chairs
Srinivas Bangalore
Christine Doran
Josef van Genabith
Hitoshi Isahara (NICT, Japan)
Philipp Koehn (University of Edinburgh, UK)
Kim-Teng Lua (COLIPS, Singapore)

Student Chairs
Davis Dimalen (Academia Sinica, Taiwan)
Jenny Rose Finkel (Stanford University, USA)
Blaise Thomson (Cambridge University, UK)

Student Workshop Faculty Advisors
Grace Ngai (Polytechnic University, Hong Kong)
Brian Roark (Oregon Health & Science University, USA)

Tutorial Chairs
Diana McCarthy (University of Sussex, UK)
Chengqing Zong (Chinese Academy of Sciences, China)

Workshop Chairs
Jimmy Lin (University of Maryland, USA)
Yuji Matsumoto (NAIST, Japan)

Webmaster
Minghui Dong (Institute for Infocomm Research, Singapore)

Registration
Priscilla Rasmussen (ACL)

Program Committee

For ACL-IJCNLP

Program Co-Chairs
Jian Su (Institute for Infocomm Research, Singapore)
Janyce Wiebe (University of Pittsburgh, USA)

Area Chairs
Eneko Agirre (University of Basque Country, Spain)
Sophia Ananiadou (University of Manchester, UK)
Anja Belz (University of Brighton, UK)
Giuseppe Carenini (University of British Columbia, Canada)
Hsin-Hsi Chen (National Taiwan University, Taiwan)
Keh-Jiann Chen (Sinica, Taiwan)
James R. Curran (University of Sydney, Australia)
Jianfeng Gao (MSR, USA)
Sanda Harabagiu (University of Texas at Dallas, USA)
Philipp Koehn (University of Edinburgh, UK)
Grzegorz Kondrak (University of Alberta, Canada)
Helen Mei-Ling Meng (Chinese University of Hong Kong, Hong Kong)
Rada Mihalcea (University of North Texas, USA)
Massimo Poesio (University of Trento, Italy)
Ellen Riloff (University of Utah, USA)
Satoshi Sekine (New York University, USA)
Noah A. Smith (CMU, USA)
Michael Strube (EML Research, Germany)
Jun Suzuki (NTT, Japan)
Haifeng Wang (Toshiba, China)

For EMNLP

Program Co-Chairs
Philipp Koehn (University of Edinburgh, UK)
Rada Mihalcea (University of North Texas, USA)

Area Chairs
Stephen Clark (University of Cambridge, UK)
Mona T. Diab (Columbia University, USA)
Jason Eisner (Johns Hopkins University, USA)
Katrin Erk (University of Texas, USA)
Eric Fosler-Lussier (Ohio State University, USA)
Iryna Gurevych (Darmstadt University, Germany)
Hang Li (Microsoft Research Asia, China)
Chin-Yew Lin (Microsoft Research Asia, China)
Adam Lopez (University of Edinburgh, UK)
Vivi Nastase (EML Research, Germany)
Miles Osborne (University of Edinburgh, UK)
Tim Paek (Microsoft, USA)
Marius Pasca (Google, USA)
Carlo Strapparava (FBK-Irst, Italy)
Theresa Wilson (University of Edinburgh, UK)

Local Arrangements Chair
Priscilla Rasmussen (ACL)

Local Committee

Local Organizing Chair
Haizhou Li (Institute for Infocomm Research, Singapore)

Webmaster and Secretariat
Minghui Dong (Institute for Infocomm Research, Singapore)

Student Volunteer Programme, Social Programme, Registration
Swee Lan See (Institute for Infocomm Research, Singapore)

Registration
Wang Xi (Institute for Infocomm Research, Singapore)

Wireless Internet
Vladimir Pervouchine (Institute for Infocomm Research, Singapore)

Posters, Exhibition and Demos
Long Qiu (Institute for Infocomm Research, Singapore)

EMNLP and Main Conference Technical Sessions
Hwee Tou Ng (National University of Singapore, Singapore)

Audio/Visual, Workshops & Collocated Events Technical Sessions
Min Zhang (Institute for Infocomm Research, Singapore)

Conference Handbook, Publicity
Min-Yen Kan (National University of Singapore, Singapore)

Printing and Publication
Eng Siong Chng (Nanyang Technological University, Singapore)

Finance
Tse Min Lua (COLIPS, Singapore)

Sponsorship Liaison
Kim-Teng Lua (COLIPS, Singapore)

Logistics, Delegate Liaison
Lawrence Por (Institute for Infocomm Research, Singapore)

Graphics
Adrian Tay (Institute for Infocomm Research, Singapore)

Blog
Chris Henry (National University of Singapore, Singapore)

Message from the General Chair

Welcome to ACL-IJCNLP 2009, the first joint conference sponsored by ACL (The Association for Computational Linguistics) and AFNLP (Asian Federation of Natural Language Processing). The idea to have a joint conference between ACL and AFNLP was first discussed at ACL-05 (Ann Arbor, Michigan) between Martha Palmer (ACL President), Benjamin T'sou (AFNLP President), Jun'ichi Tsujii (AFNLP Vice President) and Keh-Yih Su (AFNLP Conference Coordinating Committee Chair, also the Secretary General). We are glad that the original idea has come true four years later, and even the affiliation relationship between these two organizations has been built up now.

In this joint conference, we have tried to mix the spirit from both ACL and AFNLP, and Singapore, which itself has mixed cultures from various eastern and western regions, is certainly a wonderful place to see how different languages meet each other. We hope you will enjoy this big event held in this garden city, which is brought to you via the effort from each member of the conference organization team.

Among our hard working organizers, I would like to thank the Program Chairs, Jan Wiebe and Jian Su, who have carefully selected papers from our record high submissions, and the Local Arrangements Chair, Haizhou Li, who has shown his excellent capability in smoothly organizing various events and details. My thanks will also go to other chairs for their competent and hard work: The Webmaster, Minghui Dong; the Demo Chairs, Gary Geunbae Lee and Sabine Schulte im Walde; the Exhibits Chairs, Timothy Baldwin and Philipp Koehn; the Mentoring Service Chairs, Hwee Tou Ng and Florence Reeder; the Publication Chairs, Jing-Shin Chang and Regina Barzilay; the Publicity Chairs, Min-Yen Kan and Andy Way; the Sponsorship Chairs, Hitoshi Isahara and Kim-Teng Lua; the Student Research Workshop Chairs, Davis Dimalen, Jenny Rose Finkel, and Blaise Thomson; also the Faculty Advisors, Grace Ngai and Brian Roark; the Tutorial Chairs, Diana McCarthy and Chengqing Zong; the Workshop Chairs, Jimmy Lin and Yuji Matsumoto; last, the ACL Business Manager, Priscilla Rasmussen, who not only provides useful advice but also helps to contact more sponsors and get their support.

Besides, I need to express my gratitude to the Conference Coordination Committee for their valuable advice and support: in which Bonnie Dorr (Chair), Steven Bird, Graeme Hirst, Kathleen McCoy, Martha Palmer, Dragomir Radev, Priscilla Rasmussen, Mark Steedman are from ACL; and Yuji Matsumoto, Keh-Yih Su, Jun'ichi Tsujii, Benjamin T'sou, Kam-Fai Wong are from AFNLP.

Last, I sincerely thank all the authors, reviewers, presenters, invited speakers, sponsors, exhibitors, local supporting staff, and all the conference attendants. It is you that makes this conference possible. Wish you all enjoy the program that we provide.

Keh-Yih Su
ACL-IJCNLP 2009 General Chair
August 2009

Contents

1 About Singapore

For the first time, the flagship conferences of the Association of Computational Linguistics (ACL) and Asian Federation of Natural Language Processing (AFNLP) – the ACL and IJCNLP – are jointly organized as a single event. The ACL-IJCNLP 2009 will cover a broad spectrum of technical areas related to natural language and computation. You are welcome to participate in the conference to discuss the latest research findings. The 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP) will also be collocated with ACL-IJCNLP 2009 right here in Singapore.

With this exciting line-up of events, we welcome members of the various research communities to Singapore. Popularly known as "The Garden City" for its lush, green landscape couching an advanced infrastructure, the island is situated at the southern tip of the Malaysian Peninsula. This small yet properous South-East Asian nation is a cosmopolitan city brimming with diversity, which is well connected by more than 70 airlines flying to more than 160 cities in 53 countries. Natives of Singapore traditionally comprise members of the Chinese, Malay, Indian, and Eurasian ethnic groups, while its residents today hail from the world over. The multicultural heritage of Singaporeans is very much alive through the different languages, cuisines, and ethnic lifestyles. This rich cultural backdrop is certainly no hindrance to modernity and progress, as demonstrated by Singapore's icon as a shopping and dining paradise. With the many museums, theme parks and a bustling nightlife, it is indeed an excellent spot to unwind with the family or alone while enjoying the convenience of a modern city.

The conference will take place at Suntec City Singapore International Convention & Exhibition Center, which is only 20 minutes from Changi International Airport. Suntec Singapore offers direct access to 5,200 hotel rooms, 1,000 shops, 300 restaurants and a world-class performing arts center. Facilities are interconnected and easily accessible via air-conditioned tunnels and covered walkways. At no time are you more than a 15-minute walk from Suntec City.

2 Organization

ACL

The Association for Computational Linguistics is *the* international scientific and professional society for people working on problems involving natural language and computation. Membership includes (among other things) electronic issues of the journal, Computational Linguistics, reduced registration at most ACL-sponsored conferences, discounts on publications of participating publishers and ACL and related publication back issues, announcements of ACL and related conferences, workshops, and journal calls of interest to the community, and participation in ACL Special Interest Groups.

The ACL journal, *Computational Linguistics*, continues to be the primary forum for research on computational linguistics and natural language processing. Since 1988, the journal has been published for the ACL by MIT Press to provide a broader distributional base.

An annual meeting is held each summer in locations where significant computational linguistics research is carried out.

AFNLP and IJCNLP

The International Joint Conference on Natural Language Processing (IJCNLP) is the flagship conference of the Asian Federation of Natural Language Processing (AFNLP). AFNLP was founded in 2003, with the mission to promote and enhance research and development relating to the computational analysis and the automatic processing of all languages of importance to the Asian region without regard to differences in race, gender, language, religious belief or political stand, by assisting and supporting like-minded organizations and institutions through information sharing, conference organization, research and publication coordination, and other forms of support in consonance with its mission. The official members of the AFNLP are professional associations and research institutions/universities in countries or territories of the region, representing researchers in the countries/territories and undertaking the responsibility to represent them. AFNLP also welcomes other organizations – such as organizers of conferences in the region, international professional bodies, professional associations of research fields related with NLP – to join us as liaison members. Previous IJCNLP conferences were held in Sanya, China (2004), Jeju, Korea (2005), Hyderabad, India (2008).

3 Information

Instruction for Presenters

Lecture Presentation

These presentation instructions are applicable for all oral sessions in ACL-IJCNLP Main Conference, Workshops, Collocated Events, Tutorials, and EMNLP.

Equipment

Each presentation room is equipped with a laptop computer, a data projector, a microphone (for large rooms), a lectern, and a pointing device. You are strongly recommended to use the laptops provided by the conference. Identical laptops with the same specifications are also available in the Speaker Ready Room (MR204, Level 2, Suntec). You can check if your slides can be displayed properly in the Speaker Ready Room.

The laptops are equipped with:

- Windows XP SP3
- Wireless LAN connection, USB port, DVD player
- Microsoft Office 2003
- Adobe Reader, Flash Player, Media Players (Microsoft/Real/QuickTime)
- Anti-Virus software

You are advised to check if your PowerPoint slides can be displayed properly using PowerPoint Viewer 2003. WiFi is available at the conference venue. However, the bandwidth is only enough for web browsing and email, not for video/audio streaming.

Presentation

Please notify the session chair of your presence and upload your slides to the laptop in your presentation room at least 15 minutes before the start of the session. As some short paper oral sessions (including 4DII and 6E-H) are connected to the previous sessions, presenters at such sessions need be ready even before the previous session starts. A 20 minute talk time plus 5 minutes question answering are allocated for each main conference full paper and EMNLP oral presentation paper, while 12 plus 3 minutes are given per main conference short oral presentation paper. The allocated presentation time for the ACL-IJCNLP main conference, workshops, collocated events, tutorials and EMNLP may differ. Please check the conference web site for the exact time allocation for your presentation. Please rehearse your presentation and ensure it falls within the time limit. Make sure each of your key points is easy to explain with aid of the material on your slides. Do not read directly from the slide during your presentation. You should not need to prepare a written speech, although it is often a good idea to prepare the opening and closing sentences in advance.

Venue & Timetable

All technical sessions are held in Suntec Convention Center. You may check the assigned presentation room and the session timetable at the conference web site for updates.

Poster Presentation

These presentation instructions are applicable for posters in ACL-IJCNLP Main Conference Short Paper Posters, Software Demonstrations, EMNLP Poster-cum-Reception, and Student Research Workshop (A0 posters) and in ACL-IJCNLP Workshops, Collocated events (A1 posters).

A paper presented as a poster offers a unique opportunity to present a research work in a way customized to individual or a small group of people. It is more interactive than an oral presentation. Therefore, the work can be presented, from certain aspects, more effectively to a small but well-targeted audience. Remember people attracted by a poster are so interested in the work that they are willing to invest anywhere from 5 to 10 minutes of their time. That is a big chunk out of what they have for a poster session!

To attract the audience who would be interested in your work, the poster should have a title in large font which is highly visible to even passers-by. Its contents also need fonts large enough to be readable from 1 to 2 meters away. Highlight the pages of your paper in the proceedings so people can locate it easily. As to its layout, it may not be a good idea to simply create a few slides and patch them together like an enlarged handout of an oral presentation. Instead, a poster allows you to arrange things anywhere you want. For example, the system diagram can be in the center, surrounded by descriptions and performance tables of its individual components. So take advantage of this flexibility.

"A picture is worth a thousand words". Try to choose visual aids like figures, diagrams, cartoons, colors, even lines over texts on your poster to show the research idea and the logic flow of the contents. Thus after attracting people with a nice title, the poster can be self-explaining so that people can understand it and quickly find out whether they have more questions to ask. If they do, they can have a short discussion with you to get the most out of your poster presentation. In addition, some people are more verbal than visual. They prefer to listen instead of to read even the visualization is great. So, prepare "mini-talks" as short as 30 seconds, and some as long as 5 minutes. Kindly ask the people (who might be reading the poster apparently slowly) whether they would like a brief introduction from you. Similar as delivering an oral presentation, bear your audience's background in mind. Seasoned researchers probably need only a few key points while more general information would help those not so familiar with your task. If you find it will otherwise take long to talk with a highly interested person, try to wrap up politely so you can talk to other people who are waiting.

Occasionally, people prepare some printouts to complement their posters. If you expect such printouts to be helpful, please prepare them.

For posters in ACL-IJCNLP Main Conference, Short Paper Posters, Software Demonstrations, EMNLP Poster-cum-Reception, and Student Research Workshop, we will provide display easels measuring 100cm in width and 200cm in height, with a usable board area of 95cm X 190cm. This size is good for a standard A0 poster in the portrait orientation. The poster easels are double-sided with one poster on each side. Mounting items such as push tacks or Blu-Tack adhesive will be provided. However, no tables will be available except for Software Demonstrations.

For posters in ACL-IJCNLP 2009 Workshops, Collocated events, we will provide display easels in the workshop rooms that are only good for standard A1 posters in the portrait orientation. To avoid leaving your poster without a presenter (in such case it will attract less people than it deserves), try to have your co-author or colleague cover you if possible.

Internet Connectivity

Suntec City, Level 2

(Free!)

Wireless Internet access covered by our conference's WiFi signal will be available for conference participants on Level 2 of the Suntec Convention Center. The service is intended for general purpose Internet surfing, such as email, and not for high data rate usage like Internet phone (e.g., Skype), video streaming and online gaming. There is a limit of 100 concurrent clients within our conference area. The limit is distributed around Level 2 according to pre-installed access points in the building. Username and password will be provided to all participants at the registration booth.

In addition, an Internet kiosk with 10 PCs with LAN connection will be available for general purpose Internet surfing.

Wireless@SG

Internet access with wide coverage in Singapore (Free!)

Singapore offers countrywide free WiFi Internet. The service is available in selected places, like libraries, shopping malls, restaurants, etc., by connecting to WiFi network named "Wireless@SG". The full coverage information is available here. ACL-IJCNLP 2009 venue, Suntec Convention Center, has Wireless@SG coverage on the first (ground) floor.

Accessing the Internet via Wireless@SG requires username and password, which any person can obtain upon registration with one of Wireless@SG operators. The registration links are:

- Singtel – *https://myad1.singnet.com.sg/wireless@sg_signup/ onlineapplication.jsp?apptype=was*
- QMax – *http://wsg.qmax.com.sg/wreg.aspx?notify=email*
- iCell – *http://www.icellnetwork.com/reg2.php*

A few fields in the Wireless@SG registration form may require clarification:

- NRIC/FIN: Enter your passport number here.
- NRIC/FIN Type: Choose “Passport”.
- Nationality: If you cannot find yours, choose “Others”.
- Address: Enter address of the hotel you booked to stay in Singapore.
- Verification code: Type the code you see in the upper box into the lower box.

The information for registering a WirelessSG account has recently changed. Registration confirmation and password are sent via SMS; local Singapore mobile phone numbers **are now required** at the moment (International phone numbers are currently not allowed). We advise you to purchase a prepaid SIM card upon arrival to Singapore. The prepaid SIM cards for all 3 mobile operators are sold in a number of shops including convenience stores like 7-Eleven. A valid passport is required to be shown in order to purchase a SIM card. Providing a correct phone number is therefore important for successful registration. More information about Wireless@SG can be obtained from the Infocomm Development Authority website:

http://www.ida.gov.sg/Programmes/20061027174147.aspx?getPagetype=36

Starhub 3.5G HSDPA Mobile Internet Access with country wide coverage(Paid Service)

Suntec Convention Center is offering a country wide paid internet access service (wireless access) based on the rental of 3.5G HSDPA Mobile Data Card. With the rental of this service, conference participant will be able to enjoy dedicated and high bandwidth internet access everywhere in Singapore. Starhub is one of the main service providers in Singapore that offers excellent wireless coverage. Listed here are the specifications and rental costs:
Starhub 3.5G HSDPA Mobile Internet Specifications:

1. Download speed = 7.2Mbps, upload speed = 1.9MBps
2. Unlimited download usage.
3. OS supported: Windows@2000, Windows@XP, Windows@Vista and MAC OS up to 10.4 (tiger)
4. HSDPA data card provided: Huawei E170 or E270.

Starhub 3.5G HSDPA Mobile Internet Rental Cost:

1. 3 days rental inclusive of 3G SIM Card and HSDPA Data Card = 400.00 SGD
2. 7 days rental inclusive of 3G SIM Card and HSDPA Data Card = 500.00 SGD
3. Device deposit cost (refundable) = 300.00 SGD

This paid service is open for ordering now!

1. Fill in the entire form including credit card authorization portion.
2. Collection of the service package (inclusive of Data Card and SIM Card) at Suntec City. Contact person: Mr. Winston Sze, Tel: 6825 2020. Collection time varies; if you place your order prior to 15th July 2009, you can collect the package by 10:00 on 2 August (the first day of the conference). For on-site ordering:
   - before 10:00, delivery by 14:00.
   - in between 10:00 to 14:00, delivery by 18:00.
   - after 14:00, delivery by 10:00 on the next day.

Payment can be made either by credit card or by cash during collection of the package. This paid service is offered by Suntec Convention Center and therefore participants should communicate with Suntec directly.

Awards

Best Paper Awards

In the ACL and IJCNLP traditions, one or more of the conference's scientific contributors will be awarded Best Paper Awards. The conference also recognizes significant student contributions through Best Student Paper Awards. The recipients will be announced in a plenary session at the end of the conference and will receive a certificate.

Lifetime Achievement Award

The ACL Lifetime Achievement Award (LTA) was instituted on the occasion of the Association's 40^{th} anniversary meeting. The award is presented for scientific achievement, of both theoretical and applied nature, in the field of Computational Linguistics. Currently, an ACL committee nominates and selects at most one award recipient annually, considering the originality, depth, breadth, and impact of the entire body of the nominee's work in the field. The award is a crystal trophy and the recipient is invited to give a 45-minute speech on his or her view of the development of Computational Linguistics at the annual meeting of the association. As of 2004, the speech has been subsequently published in the Association's journal, *Computational Linguistics*. The speech is introduced by the announcement of the award winner, whose identity is not made public until that time.

Previous winners of the distinguished award have been: Aravind Joshi (2002), Makoto Nagao (2003), Karen Spärck Jones (2004), Martin Kay (2005), Eva Hajičová (2006), Lauri Karttunen (2007) and Yorick Wilks (2008).

Social Events

Welcome Reception – 2 August

18:00 to 21:00

The Welcome Reception is free for all ACL-IJCNLP 2009 registered participants. Please join us at Ballroom 1 for the reception to meet with your old colleagues and meet some new faces!

Student Lunch – 3 August

11:50 to 13:20

The Student Lunch is a traditional ACL event where students in the field get a chance to network with each other in a relaxed environment. It is provided free of charge to student members, but you must register in advance during the conference registration.

Please join us on the Concourse (Suntec Level 3) during lunchtime.

Banquet – 4 August

Pre Dinner Cocktail
Time: 18:30 to 19:30
Venue: Marina Mandarin Singapore, Pool Garden – Level 5

Chill out and enjoy the greenery at Marina Mandarin Pool Garden while waiting for the banquet to start. Come early for a taste of Singapore special local treats!

Banquet Dinner
Time: 19:30 to 22:00
Venue: Marina Mandarin Singapore, Ballroom – Level 1

Experience Singapore's charm and hospitality. Be sure to be treated to our mouth-watering Asian delights while enjoying stirring music. You'll certainly be tempted to groove to the beat. Look out for special "Uniquely Singapore" performances you will never forget!

To get to the banquet at Marina Mandarin from Suntec:

1. Exit Suntec International Convention and Exhibition Center via the side doors at Level 2 and take the skybridge across to Marina Square.

2. Enter Marina Square and follow the signage to the Marina Mandarin Hotel. Do not enter the hotel via Level 2. Instead, kindly take the escalator up to Level 3 and enter through the side doors.

3. Once in the Marina Mandarin hotel, take the lift up to Level 5 – Pool Garden for the pre dinner cocktail, or take the lift to Level 1 – Ballroom for the dinner.

Local Information

Emergency Phone Numbers

Singapore Country Code: 65

Police: 999 (toll-free)

Emergencies/Ambulance/Fire: 995

Non-emergency ambulance: 1777

International Calls

In addition to cheap international calls via VoIP services such as Jajah and Skype, inexpensive international calls are available from mobile operators in Singapore. Different mobile operators offer free international calls to different destinations, meaning the charges are for local airtime only. Prepaid SIM cards for Singtel, Starhub and M1 are sold in convenience stores such as 7-Eleven. A passport is required to purchase a SIM card.

Starhub – Free calls to Bangladesh, Brunei, Canada, China, Hong Kong, India, Laos, Macau, Malaysia, Puerto Rico, Russia, South Korea, Taiwan, Thailand, USA (states in USA only, including Alaska and Hawaii). Other international call rates and details are available from Starhub website.

Singtel – Free calls to Bangladesh, Brunei, Canada, China, Hong Kong, India, Malaysia, Puerto Rico, South Korea, Thailand, USA. More details available from Singtel website (http://home.singtel.com/upload_hub/consumer/faqFIDD24mar.htm). Information on other prepaid SIM cards is available from Singtel website.

Mobile One (M1) – Free calls to India, Macau, Bangladesh, Brunei, China, Vietnam, Malaysia, Australia, Thailand, Laos, Pong Kong, New Eolande, Taiwan, Peaty Rico, USA, Russia, Canada, United Kingdom, South Korea. Other international call rates and details are available from the M1 website.

Swine Flu (Influenza A – H1N1)

2009 Influenza A (H1N1), also written as Influenza A (H1N1-2009) (previously referred to as "new strain of swine flu") is a new strain of influenza virus that spreads from human to human. As this is a new strain of virus, most people will not have resistance, and it can potentially spread quickly and infect a large proportion of the population in a short period of time.

The Singapore Ministry of Health (MOH) has an official website updated with the latest information and status about Swine Flu locally:

http://www.h1n1.gov.sg/

Common symptoms of influenza includes: chills, fever, sore throat, muscle pains, headache, coughing, weakness and general discomfort.

Reducing risk of infection: Influenza viruses can be inactivated by soap, and alcohol based hand rubs. Following good hygiene practices reduces the risk of infection.

- hand washing with soap and water, or alcohol based hand rubs
- avoiding spitting
- covering the nose and mouth when sneezing or coughing, e.g. with tissue paper, and disposing the used tissue properly.

In Singapore, there are about 450 Pandemic Preparedness Clinics (PPC) available to help assess whether one is being infected with Influenza A (H1N1). You can identify these clinics from their red check mark decal with the words ‘H1N1 ready’. There are two medical clinics near the Suntec Convention Center, which are PPC.

Pharmacy

Suntec City has a few health, fitness and pharmacies on its premise. Guardian, Watsons are two large franchises that serve such needs. Many branches of these stores also have an on-duty pharmacists and prescriptions desk where you can have your medical prescriptions filled.

If you think you are developing influenza-like illness, please be socially responsible to put on a surgical mask. You can request for a surgical mask from the conference First Aider. Please visit the First Aid desk set up in the conference area. If your health condition deteriorates (e.g. fever, cough, sore throat, runny nose), please seek medical attention promptly. In case of breathing difficulty, you should seek help to call 995 immediately.

Medical Services

Clinics. There are two clinics that are also very close by if you do not require emergency medical services. You can just walk in to the clinics during regular opening hours to see a doctor. These are also the Pandemic Preparedness Clinics (PPC).

Bethesda Medical Center
#02-080 Suntec City Mall
Tel: 6337 8933
Fax: 6337 4233

Opening Hours:
Mon to Fri: 8:30 – 12:30, 14:00 – 17:00
Sat: 9:00 – 11:30
Sun: Closed

Map: *http://www.street-directory.com/hpb.ppc/index.php/map/clinics/clinic/425*

Raffles Medical
Millenia Walk
9 Raffles Boulevard,
#02-24B Millenia Walk
Tel: 6337 6000
Fax: 6334 8607

Opening Hours
Mon to Fri: 8:00 – 13:00, 14:00 – 17:30

Sat: 8:30 – 13:00 (closed on 2nd and 4th Saturdays of every month)
Sun: Closed

Map: *http://www.street-directory.com/hpb.ppc/index.php/map/clinics/clinic/390*

Hospital. The closest hospital to Suntec City is Raffles Hospital (about 1 km north), at 585 North Bridge Road. It is right next to the Bugis MRT station on the East West Line). It provides 24-hour emergency services, dental clinic, traditional Chinese medicine (TCM) and also specializes in other outpatient services for international clients.

The 24-Hour Emergency phone number to call for ambulance services or specialist on call at the Raffles Hospital is 6311 1555.

Post Office

There is a Singapore Post branch to service all of your mailing needs, located within Suntec City. The post office is located on the third floor of the mall.

Suntec City
3 Temasek Boulevard #03-001/003
Suntec City Mall
Singapore 038983
Opening hours: Mon-Fri: 9:30 to 18:00, Sat: 9:30 to 14:00
Closed on Sundays and public holidays.
Telephone: 6332 0289

Shopping

Suntec City, Citylink and the Raffles City shopping malls will be able to meet most of your shopping needs. For more shopping, proceed up Orchard Road and its line of shopping plazas. For electronic buffs, Funan IT Mall and Sim Lim Plaza carry a large selection of electronic goods. Most malls in Singapore serve the public starting from 10:00 till 22:00, daily.

To get your shopping fix 24/7, the well-known Mustafa Center is open all day, every day, and is located at the Northern end of the Little India.

For a more regional flavor, try going through Chinatown, Little India or Kampong Glam, for a taste of the three largest ethnic groups' indigenous cultural flavor.

- *Sim Lim Square* 1 Rochor Canal Road, Singapore 188504. Tel: 6338 3859, Fax : 6334 3469. Closest MRT Station: Bugis (EW12) and Little India (NE7). Operating Hours: 10:30 – 21:00 daily.

- *Mustafa Center*, 145 Syed Alwi Road, Singapore 207704 Tel: 6295 5855, Fax: 6295 5866. Closest MRT Station: Farrer Park.

- *Funan IT Mall*, 109 North Bridge Road, Singapore 179097 Tel: 6336 8327. Closest MRT Station: City Hall. Operating Hours: 10:00–22:00 daily.

- *Chinatown*. Closest MRT Station: Outram Park (EW16/NE3) or Chinatown (NE4).

- *Little India*. Closest MRT Station: Little India (NE7) or take SBS Transit bus number 65 from Orchard Road, alight at Tekka Market along Serangoon Road.

- *Kampong Glam*. Closest MRT Station: Bugis (EW12) and walk along Victoria Street towards Arab Street and Sultan Gate Street towards the Malay Heritage Center.

Eating

You have a variety of choices for eating in Suntec City. Most food courts or hawker centers will have vegetarian and Halal selections. Diners with Kosher requirements have few options, but if you can eat fruit and/or vegetarian meals there will be a wider variety of choices.

A full list of dining options in Suntec can be found here:

http://www.sunteccity.com.sg/chinese/retail/western.htm

For a quick bite, we recommend the Food Republic downstairs from the conference venue.

Food Republic
Suntec City, Level 1
Mon – Thu: 10am – 10pm
Fri-Sun and Public Holidays: 10:00 – 23:00

Set amidst an "Old Europe" ambiance, Food Republic is prominently located on the first level of the convention center. Its decor is conceptualized as a gentleman's nineteenth-century library. Designer wallpaper, antique bookends and chandeliers are among the numerous touches that give the feel of old world glamour. Even the food is served on fine China. Completing the ambiance is the service staff whose attire calls to mind Victorian elegance.

The 14 food stalls were handpicked for their heritage recipes, authentic flavors and reputation among local diners and food critics. Diners can look forward to local favorites such as Bak Kut Teh, Hainanese Chicken Rice, the famed JB Ah Koong fish balls, Muslim favorites, Roti Prata and even Dim Sum and Japanese fare.

If this is still not enough, you can also join a Peranakan trail from the conference travel agency:

http://www.globaltravel.com.sg/eflash/html/singaporesightseeingtours01.html

Electricity

Singapore uses the UK standard for electricity and electrical plugs: 230V/50Hz – British plug.

Smoking

Singapore bans smoking in most public spaces, even some outdoor dining spaces. Smokers are advised to check with the establishment whether there are designated seats where smoking is permitted. Many shopping areas also have a designated outdoor smoking area.

Transportation

Suntec City is centrally located within the commercial district of Singapore. Walking will get you to many places of interest within a kilometer of the conference venue. For farther trips, you may want to consider taking the public buses (from various bus shelters located at regular intervals along on the roads) or the Mass Rapid Transit (MRT) rail lines. The nearest MRT station is at City Hall (walk from Suntec through the CityLink mall to reach the station), about a 10 minute walk. Bus and MRT fares typically run between 50 cents to 2 dollars. The East West line of the MRT system runs directly to Changi Airport (a same platform transfer at Tanah Merah is needed), with total travel time originating at the City Hall MRT, taking about 45 minutes.

For more convenience, taxis are convenient to take. Note that around the Suntec City area, taxis are required to pick up and drop off passengers only a designated taxi stands. The closest taxi stand is right at the entrance of the conference venue, the Suntec City Convention Center. All hotels also serve as taxi stands. Fares often start at 3 SGD and for journeys around the downtown area should cost no more than 10 SGD. There is a surcharge for morning rush, evening rush and after midnight periods and if road tolls (ERP) are crossed.

Taxis from the city area to the airport are available at all hours of the day and advanced booking is usually not needed, except if you expect to run into peak period traffic. A fare to the airport should run under 20 dollars for most. Consult your hotel or hostel to see whether advanced booking is recommended.

CityCab Taxi:	6552 2222
Comfort CabLink Taxi:	6552 1111
TIBS Taxi:	6555 8888
Yellow-Top Cab:	6552 2828

Airport Flight Information: 1 800 542 4422

National Day – Sunday, 9 August

At Marina Bay

Singapore celebrated its first National Day in 1966, one year after Singapore's independence from Malaysia on 9 August 1965.

Over the years, the National Day Parade (NDP) has become the biggest national event in Singapore. What is perhaps most memorable at each celebration is the fireworks display marking the climax of the parade; the sky would be bursting with the wonderful colors of the visual vista, dazzling it as well as the hearts of fellow Singaporeans. On this very special occasion, most Singaporeans would be decked out in patriotic colours – namely, red and white.

The Parade has gained enormous popularity and support from the people that it is not unusual to find massive number of citizens trying to get their hands on a ticket, which is released free-of-charge. People would arrive hours before the ticket booths opened to release tickets. This proved to be problematic and as such, the government set up the e-balloting ticketing system in 2003.

Traditionally held at either on the Padang or the National Stadium, the 2009 event will be held at the Marina Bay Floating Stadium this year. While this means it is very difficult even for citizens to see the parade seated in the Stadium (only 30,000 seats), there are estimated 150,000 expected people who will see the NDP from vantage points around the Marina Bay area, of which Suntec borders. You'll be able to see and hear the festivities of NDP 2009 from wherever you are in Singapore on the 9th! We're sure you'll enjoy the occasion!

Other Frequently Asked Questions

1. Any ACL-IJCNLP 2009 souvenirs?

   ACL-IJCNLP 2009 souvenirs (Limited Edition) are available at the registration desk.

   - ACL-IJCNLP 2009 conference T-shirt
   - Special edition of Singapore stamps celebrating ACL-IJCNLP 2009
   - Photos taken by the photographers invited by ACL-IJCNLP 2009 (free of charge)

2. What's the current exchange rate for SGD in various major currencies?

   - 1 U.S. Dollar = 1.5 SGD
   - 1 Euro = 2 SGD
   - 1 British Pound = 2.4 SGD
   - 100 Japanese Yen = 1.5 SGD
   - 10 Chinese Yuan = 2.1 SGD

3. Where can I find a money changer?

   There are banks, ATMs, and money changers located in the Entertainment Center, Tropics, Fountain Terraces of the Suntec City Mall, just next door to the Suntec Convention Center. You can check out their exact locations from the Suntec City Website (*http://www.suntecreit.com/sunteccitymall/home.aspx*). If you are still lost, or need more directional assistance around Suntec City or Singapore, you can always approach the local conference volunteers for their assistance or advice.

4. What if I feel unwell, or need first aid assistance during the conference?

   Please visit the First Aid desk in the conference area immediately, or approach any of the conference volunteers for their assistance. There are First Aiders in the conference, who can assist you. If you think you are developing influenza-like illness, and want to be socially responsible, you can request for a 3-ply surgical mask to use. In case you need to visit a clinic, there are conference officers who can also help escort you to the nearby clinic, or help you call 995 for the ambulance service. Outside the conference hours, you can refer to this handbook for self-assistance, or contact your hotel front desk for their services.

Sponsors

The ACL-IJCNLP 2009 very gratefully acknowledges the following committments in sponsorship:

ACL-IJCNLP 2009 is proudly co-organized by COLIPS, Singapore.

COLIPS
celebrating 21 years

The Chinese and Oriental Languages Information Processing Society, or COLIPS in short, is a non-profit professional organisation that was established in 1988 to advance the science and technology of information processing in Chinese and other Asian languages. It promotes the free exchange of information about information processing of these languages in the best scientific and professional tradition. COLIPS organizes international conferences, short courses and seminars for members and the public. It is one of the founding members of Asian Federation of Natural Language Processing (AFNLP). COLIPS publishes the International Journal of Asian Languages Processing four times a year that is circulated world-wide. Having its members from all over the world, COLIPS is based in Singapore. In 2009, COLIPS proudly organises the first joint conference between ACL and AFNLP, ACL-IJCNLP 2009, in Singapore while it celebrates its 21st anniversary.

www.colips.org

ACL-IJCNLP 2009 is proudly co-organized by Institute for Infocomm Research, Singapore.

A*STAR

Institute for
Infocomm Research

The Institute for Infocomm Research (I^{2}R pronounced as "I-squared-R") is a member of the Agency for Science, Technology and Research (A*STAR) family. Established in 2002, our mission is to be the globally preferred source of innovations in "Interactive Secured Information, Content and Services Anytime Anywhere" through research by passionate people dedicated to Singapore's economic success. I^{2}R performs R&D in information, communications and media (ICM) technologies to develop holistic solutions across the ICM value chain. Our research capabilities are in information technology, wireless and optical communication networks, interactive and digital media; signal processing and computing. We seek to be the infocomm and media value creator that keeps Singapore ahead.

www.i2r.a-star.edu.sg

CNGL
Centre for Next Generation Localisation

TOSHIBA
Leading Innovation

YAHOO!
LABS

NANYANG
TECHNOLOGICAL
UNIVERSITY

School of Computer Engineering

Speech and Language Technology Program
School of Computer Engineering
Nanyang Technological University, Singapore

Congratulate ACL and AFNLP
on ACL-IJCNLP 2009!

NUS
National University
of Singapore

School *of* Computing

Lab for Media Search – *lms.comp.nus.edu.sg*
Natural Language Processing Group – *nlp.comp.nus.edu.sg*
and WING – *wing.comp.nus.edu.sg*

Local Sponsors of the ACL-IJCNLP 2009 Conference

WHOLETREE

Supporter

Xerox Research Centre Europe

Student Travel Sponsors

National Science Foundation

The Donald and Betty Walker Student Scholarship Fund

AFNLP-Nagao Fund

COLIPS Fund

Lee Foundation

4 Conference Schedules

Daily, 2–7 August

7:30-17:30	Registration	Level 2

Sunday, 2 August

8:30-12:00	Tutorials - Morning Sessions	Level 2
14:00-17:30	Tutorials - Afternoon Sessions	Level 2
8:30-18:00	Collocated Events	Level 2
18:00-21:00	Welcome Reception	Ballroom 1

Monday, 3 August

8:30-18:00	ACL-IJCNLP Technical Program	Level 2
11:50-13:20	Student Lunch	Concourse, Level 3
17:35-19:00	Software Demos	Ballroom Foyer

Tuesday, 4 August

8:30-18:00	ACL-IJCNLP Technical Program	Level 2
10:15-12:20	Student Research Workshop (oral)	MR209
12:20-14:20	Student Research Workshop (poster)	Ballroom Foyer Level 2
12:20-14:20	Short Paper / Posters	Ballroom Foyer Level 2
18:30-19:30	Pre-dinner Cocktail	Marina Mandarin Hotel Level 5
19:30-23:00	Banquet	Marina Mandarin Hotel Level 1

Wednesday, 5 August

8:30-18:00	ACL-IJCNLP Technical Program	Level 2
12:30-14:00	ACL Business Meeting	Ballroom 2
16:40-18:45	LTA & Closing	Ballroom 2

Thursday–Friday, 6–7 August

8:30-10:00	Workshops - Morning Session 1	Level 3
10:00-10:30	Workshops - Coffee Break	Level 3
10:30-12:10	Workshops - Morning Session 2	Level 3
13:50-15:30	Workshops - Afternoon Session 1	Level 3
15:30-16:00	Workshops - Coffee Break	Level 3
16:00-18:00	Workshops - Afternoon Session 2	Level 3
8:30-10:00	EMNLP - Morning Session 1	Level 2
10:00-10:30	EMNLP - Coffee Break	Level 2
10:30-12:10	EMNLP - Morning Session 2	Level 2
13:50-15:30	EMNLP - Afternoon Session 1	Level 2
15:30-16:00	EMNLP - Coffee Break	Level 2
16:00-18:00	EMNLP - Afternoon Session 2	Level 2

Thursday, 6 August

18:00-20:00	EMNLP Poster-cum-Reception	Pre-Function**

**: Meeting Room Pre-Function Area Level 2

5

Sunday, 2 August

Venue	MR202	MR208	MR209	MR210	MR206	MR207
8:30-10:00	T1	T2	T3		TCAST	MALINDO
10:00-10:30	Coffee/Tea Break					
10:30-12:00	T1	T2	T3		TCAST	MALINDO
12:00-14:00	Lunch					
14:00-15:30	T4	T5	T6		TCAST	MALINDO
15:30-16:00	Coffee/Tea Break					
16:00-17:30	T4	T5	T6		TCAST	MALINDO
17:30-18:00	Break					
17:00-17:45				Press Conference		
18:00-21:00	Registration and Welcome Reception (Ballroom 1)					

Tutorials

Morning

T1: Fundamentals of Chinese Language Processing
Chu-Ren Huang and Qin Lu
MR202

T2: Topics in Statistical Machine Translation
Kevin Knight and Philipp Koehn
MR208

T3: Semantic Role Labeling: Past, Present and Future
Lluís Màrquez
MR209

Afternoon

T4: Computational Modeling of Human Language Acquisition
Afra Alishahi
MR202

T5: Learning to Rank
Hang Li
MR208

T6: State-of-the-art NLP approaches to coreference resolution: theory and practical recipes
Simone Paolo Ponzetto and Massimo Poesio
MR209

Tutorial Descriptions

T1: Fundamentals of Chinese Language Processing

This tutorial gives an introduction to the fundamentals of Chinese language processing for text processing. Computer processing of Chinese text requires the understanding of both the language itself and the technology to handle them. The tutorial contains two parts. The first part overviews the grammar of the Chinese language from a language processing perspective based on naturally occurring data. Real examples of actual language use are illustrated based on a data driven and corpus based approach so that its links to computational linguist approaches for computer processing are naturally bridged in. A number of important Chinese NLP resources are presented. The second part overviews Chinese specific processing issues and corresponding computational technologies. The tutorial focuses on Chinese word segmentation with a brief introduction to Part-of-Speech tagging and some Chinese NLP applications. Word segmentation problem has to deal with some Chinese language unique problems such as unknown word detection and named entity recognition which will be the emphasis of this tutorial.

This tutorial is targeted for both Chinese linguists who are interested in computational linguistics and computer scientists who are interested in research on processing Chinese. More specifically, the expected audience comes from three groups: (1) The linguistic community - for any linguist or language scientist whose typological, comparative, or theoretical research requires understanding of the Chinese grammar and processing of Chinese text through observations to corpus data. It is also helpful for Chinese linguists, from graduate students to experts who may have good knowledge of the language to learn methods to process Chinese text data using computational means; (2) For researchers and students in computer science who are interested in doing research and development in language technology for the Chinese language; (3) For scholars in neighboring fields who work on Chinese, such as in communication, language learning and teaching technology, psychology, and sociology: for description of basic linguistic facts, and as resources for basic data.

Some basic knowledge of Chinese would be helpful. Comprehensive understanding of the language is not necessary.

Presenters:

Chu-Ren Huang
Dean of Humanities, The Hong Kong Polytechnic University
Research Fellow, Academia Sinica
churen.huang@inet.polyu.edu.hk

Qin Lu
Department of Computing,
The Hong Kong Polytechnic University
Hung Hom, Hong Kong
csluqin@comp.polyu.edu.hk

T2: Topics in Statistical Machine Translation

In the past, we presented tutorials called “Introduction to Statistical Machine Translation”, aimed at people who know little or nothing about the field and want to get acquainted with the basic concepts. This tutorial, by contrast, goes more deeply into selected topics of intense current interest. We envision two types of participants:

1) People who understand the basic idea of statistical machine translation and want to get a survey of hot-topic current research, in terms that they can understand.

2) People associated with statistical machine translation work, who have not had time to study the most current topics in depth.

We fill the gap between the introductory tutorials that have gone before and the detailed scientific papers presented at ACL sessions.

Presenters:

Kevin Knight
Address: 4676 Admiralty Way, Marina del Rey, CA, 90292, USA
Email: knight@isi.edu

Philipp Koehn
Address: 10 Crichton Road, Edinburgh, EH8-9AB
Email: pkoehn@inf.ed.ac.uk

T3: Semantic Role Labeling: Past, Present and Future

Semantic Role Labeling (SRL) consists of detecting basic event structures such as “who” did “what” to “whom”, “when” and “where”. The identification of such event frames holds potential for significant impact in many NLP applications, such as Information Extraction, Question Answering, Summarization and Machine Translation among others. The work on SRL has included a broad spectrum of supervised probabilistic and machine learning approaches, presenting significant advances in many directions over the last several years. However, despite all the efforts and the considerable degree of maturity of the SRL technology, the use of SRL systems in real-world applications has so far been limited and, certainly, below the initial expectations. This fact has to do with the weaknesses and limitations of current systems, which have been highlighted by many of the evaluation exercises and keep unresolved for a few years.

This tutorial has two differentiated parts. In the first one, the state-of-the-art on SRL will be overviewed, including: main techniques applied, existing systems, and lessons learned from the evaluation exercises. This part will include a critical review of current problems and the identification of the main challenges for the future. The second part is devoted to the lines of research oriented to overcome current limitations. This part will include an analysis of the relation between syntax and SRL, the development of joint systems for integrated syntactic-semantic analysis, generalization across corpora, and engineering of truly semantic features.

Presenters:

Lluís Màrquez
TALP Research Center
Software Department
Technical University of Catalonia
e-mail: lluism@lsi.upc.edu
URL: http://www.lsi.upc.edu/~lluism

T4: Computational Modeling of Human Language Acquisition

The nature and amount of information needed for learning a natural language, and the underlying mechanisms involved in this process, are the subject of much debate: is it possible to learn a language from usage data only, or some sort of innate knowledge and/or bias is needed to boost the process? This is a topic of interest to (psycho)linguists who study human language acquisition, as well as computational linguists who develop the knowledge sources necessary for large-scale natural language processing systems. Children are a source of inspiration for any such study of language learnability. They learn language with ease, and their acquired knowledge of language is flexible and robust.

Human language acquisition has been studied for centuries, but using computational modeling for such studies is a relatively recent trend. However, computational approaches to language learning have become increasing popular, mainly due to the advances in developing machine learning techniques, and the availability of vast collections of experimental data on child language learning and child-adult interaction.

Many of the existing computational models attempt to study the complex task of learning a language under the cognitive plausibility criteria (such as memory and processing limitations that humans face), as well as to explain the developmental patterns observed in children. Such computational studies can provide insight into the plausible mechanisms involved in human language acquisition, and be a source of inspiration for developing better language models and techniques.

This tutorial will review the main research questions that the researchers in the field of computational language acquisition are concerned with, as well as the common approaches and techniques in developing these models. Computational modeling has been vastly applied to different domains of language acquisition, including word segmentation and phonology, morphology, syntax, semantics and discourse. However, due to time restrictions, the focus of the tutorial will be on the acquisition of word meaning, syntax, and the link between syntax and semantics.

Presenter:

Afra Alishahi
Computational Psycholinguistics Group,
Department of Computational Linguistics and Phonetics,
Saarland University, Germany
afra@coli.uni-sb.de

T5: Learning to Rank

In this tutorial I will introduce ‘learning to rank’, a machine learning technology on constructing a model for ranking objects using training data. I will first explain the problem formulation of learning to rank, and relations between learning to rank and the other learning tasks. I will then describe learning to rank methods developed in recent years, including pointwise, pairwise, and listwise approaches. I will then give an introduction to the theoretical work on learning to rank and the applications of learning to rank. Finally, I will show some future directions of research on learning to rank. The goal of this tutorial is to give the audience a comprehensive survey to the technology and stimulate more research on the technology and application of the technology to natural language processing.

Learning to rank has been successfully applied to information retrieval and is potentially useful for natural language processing as well. In fact many NLP tasks can be formalized as ranking problems and NLP technologies may be significantly improved by using learning to rank techniques. These include question answering, summarization, and machine translation. For example, in machine translation, given a sentence in the source language, we are to translate it to a sentence in the target language. Usually there are multiple possible translations and it would be better to sort the possible translations in descending order of their likelihood and output the sorted results. Learning to rank can be employed in the task.

Presenter:

Hang Li
Microsoft Research Asia

Email: hangli@microsoft.com
Homepage: http://research.microsoft.com/en-us/people/hangli/

T6: State-of-the-art NLP approaches to coreference resolution: theory and practical recipes

The identification of different nominal phrases in a discourse as used to refer to the same (discourse) entity is essential for achieving robust natural language understanding (NLU). The importance of this task is directly amplified by the field of Natural Language Processing (NLP) currently moving towards high-level linguistic tasks requiring NLU capabilities such as e.g. recognizing textual entailment. This tutorial aims at providing the NLP community with a gentle introduction to the task of coreference resolution from both a theoretical and an application-oriented perspective. Its main purposes are: (1) to introduce a general audience of NLP researchers to the core ideas underlying state-of-the-art computational models of coreference; (2) to provide that same audience with an overview of NLP applications which can benefit from coreference information.

Presenters:

Simone Paolo Ponzetto
Assistant Professor
niversity of Heidelberg, Germany

Massimo Poesio
Chair in Humanities Computing at the University of Trento
Director of the Language Interaction and Computation Lab
Center for Mind / Brain Sciences

The Second Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST 2009)

Venue: MR206
Chairs: Linshan Lee, Haizhou Li, Luong Chi Mai, Satoshi Nakamura, Hammam Riza, Eiichiro Sumita, Chai Wutiwiwatchai

08:30 – 09:00	Registration
09:00 – 09:40	Opening Address: Asian Speech-to-Speech Translation Research Consortium - Towards Connecting Speech Translation Systems in the Asian Region *Satoshi Nakamura, Jun Park, Chai Wutiwiwatchai, Hammam Riza, Karunesh Arora, Chi Mai Luong and Haizhou Li*
09:40 – 10:00	An Overview of Korean-English Speech-to-Speech Translation System *Ilbin Lee, Jun Park, Changhyun Kim, Youngik Kim and Sanghun Kim*
10:00 – 10:30	Coffee break
10:30 – 10:50	Improvement Issues in English-Thai Speech Translation *Chai Wutiwiwatchai, Thepchai Supnithi, Peerachet Porkaew and Nattanun Thatphithakkul*
10:50 – 11:10	Toward Asian Speech Translation: The Development of Speech and Text Corpora for Vietnamese language *Thang Tat Vu, Khanh Tang Nguyen, Le Thanh Ha, Mai Chi Luong and Satoshi Nakamura*
11:10 – 11:30	Adapting Chinese Word Segmentation for Translation by Using a Bilingual Dictionary *Hailong Cao, Masao Utiyama and Eiichiro Sumita*
11:30 – 11:50	NICT/ATR Asian Spoken Language Translation System for Multi-Party Travel Conversation *Sakriani Sakti, Tat Thang Vu, Andrew Finch, Michael Paul, Ranniery Maia, Shinsuke Sakai, Teruaki Hayashi, Shigeki Matsuda, Noriyuki Kimura, Yutaka Ashikari, Eiichiro Sumita and Satoshi Nakamura*
11:50 – 13:50	Lunch

13:50 – 14:10	Development of HMM-based Hindi Speech Synthesis System *Sunita Arora, Rajat Mathur, Karunesh Arora and S.S. Agrawal*
14:10 – 14:30	Malay Multi-word Expression Translation *Aiti Aw, Sharifah Mahani Aljunied and Haizhou Li*
14:30 – 14:50	Development of Database for Speech Synthesizer In Hindi Language Using Festvox *Archana Balyan, S.S. Agrawal and Amita Dev*
14:50 – 15:10	Building a Pronunciation Dictionary for Indonesian Speech Recognition System *Amalia Zahra, Sadar Baskoro and Mirna Adriani*
15:10 – 15:30	Advances in Speech Recognition and Translation for Bahasa Indonesia *Hammam Riza and Oskar Riandi*
15:30 – 16:00	Coffee Break
16:00 – 16:20	Piramid: Bahasa Indonesia and Bahasa Malaysia Translation System Enhanced through Comparable Corpora *Aiti Aw, Sharifah Mahani Aljunied, LianHau Lee and Haizhou Li*
16:20 – 16:40	A Feature-rich Supervised Word Alignment Model for Phrase-based SMT *Chooi Ling Goh and Eiichiro Sumita*

The Third International Workshop on Malay and Indonesian Language Engineering (MALINDO 2009)

Venue: MR207
Chairs: Yussof Zaharin, Christian Boitet, Bali Ranaivo-Malançon, Mirna Adriani, Stéphane Bressan

8:00 – 8:30	Registration and Welcome

Session 1: Speech Processing

8:30 – 9:00	Malay Grapheme to Phoneme Tool for Automatic Speech Recognition *Tien-Ping Tan and Bali Ranaivo-Malançon*
9:00 – 9:20	Segmenting Indonesian Speech Documents Using Text Tiling Method *Edison Pardengganan Siahaan and Mirna Adriani*
9:20 – 9:40	The Performance of Speech Recognition System for Bahasa Indonesia Using Various Speech Corpus *Amalia Zahra, Sadar Baskoro and Mirna Adriani*
9:40 – 10:00	Development of Indonesian Spoken Language Technologies for Multilingual Speech-to-Speech Translation System *Sakriani Sakti, Michael Paul, Ranniery Maia, Noriyuki Kimura, Eiichiro Sumita and Satoshi Nakamura*

10:00 – 10:30	AM Coffee/Tea

Session 2: Information Retrieval and Applications

10:30 – 10:50	Semi-supervised Classification of Indonesian documents using the Naïve Bayes and Expectation Maximization Algorithm *Bayu Distiawan and Ruli Manurung*
10:50 – 11:10	Clustering Indonesian Document Using Non-negative Matrix Factorization and Random Projection *Suryanto Ang and Ruli Manurung*
11:10 – 11:30	Hidden Markov Model for Sentence Compression of Indonesian Language *Yudi Wibisono and Dwi Widyantoro*
11:30 – 11:50	Advance Learning and Processing Indonesian using Role Playing Game *Jasson Prestiliano and Eko Sediyono*
11:50 – 12:00	Discussion

12:00 – 13:30	Lunch

Session 3: Machine Translation

13:50 – 14:10	Poor Man's Word-Segmentation: Unsupervised Morphological Analysis for Indonesian *Harald Hammarström*
14:10 – 14:30	Tokenization of a Malay/Indonesian Corpus for Use in a Phrase-based SMT *Chooi Ling Goh and Eiichiro Sumita*
14:30 – 14:50	Applying Analogy Method to Example-Based Machine Translation (EBMT) Based on Synchronous Structured String-Tree Correspondence (S-SSTC) *Tang Enya Kong and Lim Huan Ngee*
14:50 – 15:10	Evaluating Various Corpora for Building Indonesian-English Statistical Machine Translation System *Aurora Marsye Maramis and Mirna Adriani*
15:10 – 15:30	Automatic Indonesian-English Cross-Lingual Sentence Alignment *Yunika Sugianto and Stéphane Bressan*
15:30 – 15:50	Towards Indonesian English Machine Translation : Cross Lingual News Articles Alignment *Hartanto Andreas and Stéphane Bressan*
15:50 – 16:00	Discussion
15:30 – 16:00	PM Coffee/Tea

Session 4: Short Papers

16:00 – 16:15	Structural and functional mismatches in Indonesian ber-constructions: issues in linguistic analysis and computational implementation *Wayan Arkan*
16:15 – 16:30	Probabilistic Parsing for Indonesian language *Rosa Sukamto and Dwi Widyantoro*
16:30 – 16:45	Statistical Based Part Of Speech Tagger for Bahasa Indonesia *Mirna Adriani, Hisar M Manurung and Femphy Pisceldo*
16:45 – 17:00	Automatic Tag Generation Based on Context Analysis *Putu Wuri Handayani, Made I Made Wiryana and Jan-Torsten Milde*
17:00 – 17:15	Named Entity Recognition for Indonesian *Gunawan and Dyan Indahrani*
17:15 – 17:30	Developing the Standard of the Indonesian Legal Document Using Information Extraction System *Mulyandra Mulyandra and Indra Budi*
17:30 – 17:45	Developing Indonesian Pronunciation Dictionary *Myrna Laksman-Huntley and Mirna Adriani*
17:45 – 18:00	Evaluation Statistical Machine Translation *Chairil Chairil Hakim*
18:00 – 18:15	Software Review For Young Executives in Multinational Corporation in Malaysia: The Case of Computer-Aided Writing Workbench for Young Executives *Norwati Md Yusof, Nur Ehsan Mohd Said and Saadiyah Darus*
18:15 – 18:30	Discussion and Conclusion

6

Monday, 3 August

Session	Plenary / Oral Session A	Oral Session B	Oral Session C	Oral Session D
Venue	Ballroom 2	Ballroom 1	MR 203	MR 209
8:30-8:40	Opening			
8:40-9:40	Invited Talk			
9:40-10:10	AM Coffee/Tea Break (Ballroom Foyer)			
10:10-11:50	1A: Semantics	1B: Syntax and Parsing 1	1C: Statistical and Machine Leanring Methods	1D: Phonology and Morphology
11:50-13:20	Lunch Student Luncheon (Concourse)			
13:20-15:00	2A: Machine Translation 1	2B: Generation and Summarization 1	2C: Sentiment Analysis 1	2D: Language Resources
15:00-15:30	PM Coffee/Tea Break (Ballroom Foyer)			
15:30-17:35	3A: Machine Translation 2	3B: Syntax and Parsing 2	3C: Information Extraction 1	3D: Semantics 2
17:35-19:00	Demos (11) (Ballroom Foyer)			

Invited Talk

Monday, 3 August
8:40 – 9:40, Ballroom 2

Qiang Yang, Hong Kong University of Science and Technology
Heterogeneous Transfer Learning with Real-world Applications

In many real-world machine learning and data mining applications, we often face the problem where the training data are scarce in the feature space of interest, but much data are available in other feature spaces. Many existing learning techniques cannot make use of these auxiliary data, because these algorithms are based on the assumption that the training and test data must come from the same distribution and feature spaces. When this assumption does not hold, we have to seek novel techniques for “transferring” the knowledge from one feature space to another. In this talk, I will present our recent works on heterogeneous transfer learning. I will describe how to identify the common parts of different feature spaces and learn a bridge between them to improve the learning performance in target task domains. I will also present several interesting applications of heterogeneous transfer learning, such as image clustering and classification, cross-domain classification and collaborative filtering.

Qiang Yang is a professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. His research interests are artificial intelligence, including automated planning, machine learning and data mining. He graduated from Peking University in 1982 with BSc. in Astrophysics, and obtained his MSc. degrees in Astrophysics and Computer Science from the University of Maryland, College Park in 1985 and 1987, respectively. He obtained his PhD in Computer Science from the University of Maryland, College Park in 1989. He was an assistant/associate professor at the University of Waterloo between 1989 and 1995, and a professor and NSERC Industrial Research Chair at Simon Fraser University in Canada from 1995 to 2001. Qiang Yang has been active in research on artificial intelligence planning, machine learning and data mining. His research teams won the 2004 and 2005 ACM KDDCUP international competitions on data mining. He has been on several editorial boards of international journals, including IEEE Intelligent Systems, IEEE Transactions on Knowledge and Data Engineering and Web Intelligence. He has been an organizer for several international conferences in AI and data mining, including being the conference co-chair for ACM IUI 2010 and ICCBR 2001, program co-chair for PRICAI 2006 and PAKDD 2007, workshop chair for ACM KDD 2007, AAAI tutorial chair for AAAI 2005 and 2006, data mining contest chair for IEEE ICDM 2007 and 2009, and vice chair for ICDM 2006 and CIKM 2009. He is a fellow of IEEE and a member of AAAI and ACM. His home page is at *http://www.cse.ust.hk/∼qyang*

ACL-IJCNLP – Day 1

8:30 – 8:40	Opening Session
8:40 – 9:40	Invited Talk: Heterogeneous Transfer Learning with Real-world Applications (Ballroom 2) *Qiang Yang*
9:40 – 10:10	Break

Session 1A (Ballroom 2): Semantics 1
Session Chair: *Graeme Hirst*

10:10 – 10:35	Investigations on Word Senses and Word Usages *Katrin Erk, Diana McCarthy and Nicholas Gaylord*
10:35 – 11:00	A Comparative Study on Generalization of Semantic Roles in FrameNet *Yuichiroh Matsubayashi, Naoaki Okazaki and Jun'ichi Tsujii*
11:00 – 11:25	Unsupervised Argument Identification for Semantic Role Labeling *Omri Abend, Roi Reichart and Ari Rappoport*
11:25 – 11:50	Brutus: A Semantic Role Labeling System Incorporating CCG, CFG, and Dependency Features *Stephen Boxwell, Dennis Mehay and Chris Brew*

Session 1B (Ballroom 1): Syntax and Parsing 1
Session Chair: *Christopher D. Manning*

10:10 – 10:35	Exploiting Heterogeneous Treebanks for Parsing *Zheng-Yu Niu, Haifeng Wang and Hua Wu*
10:35 – 11:00	Cross Language Dependency Parsing using a Bilingual Lexicon *Hai Zhao, Yan Song, Chunyu Kit and Guodong Zhou*
11:00 – 11:25	Topological Field Parsing of German *Jackie Chi Kit Cheung and Gerald Penn*
11:25 – 11:50	Unsupervised Multilingual Grammar Induction *Benjamin Snyder, Tahira Naseem and Regina Barzilay*

Session 1C (MR203): Statistical and Machine Learning Methods 1
Session Chair: *Jun Suzuki*

10:10 – 10:35	Reinforcement Learning for Mapping Instructions to Actions *S.R.K. Branavan, Harr Chen, Luke Zettlemoyer and Regina Barzilay*
10:35 – 11:00	Learning Semantic Correspondences with Less Supervision *Percy Liang, Michael Jordan and Dan Klein*
11:00 – 11:25	Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling *Daichi Mochihashi, Takeshi Yamada and Naonori Ueda*
11:25 – 11:50	Knowing the Unseen: Estimating Vocabulary Size over Unseen Samples *Suma Bhat and Richard Sproat*

Session 1D (MR209): Phonology and Morphology
Session Chair: *Jason Eisner*

10:10 – 10:35	A Ranking Approach to Stress Prediction for Letter-to-Phoneme Conversion *Qing Dou, Shane Bergsma, Sittichai Jiampojamarn and Grzegorz Kondrak*
10:35 – 11:00	Reducing the Annotation Effort for Letter-to-Phoneme Conversion *Kenneth Dwyer and Grzegorz Kondrak*
11:00 – 11:25	Transliteration Alignment *Vladimir Pervouchine, Haizhou Li and Bo Lin*
11:25 – 11:50	Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike *Bart Jongejan and Hercules Dalianis*

11:50 – 13:20	Lunch
11:50 – 13:20	Student Luncheon (Concourse)

Session 2A (Ballroom 2): Machine Translation 1

Session Chair: *Qun Liu*

13:20 – 13:45	Revisiting Pivot Language Approach for Machine Translation *Hua Wu and Haifeng Wang*
13:45 – 14:10	Efficient Minimum Error Rate Training and Minimum Bayes-Risk Decoding for Translation Hypergraphs and Lattices *Shankar Kumar, Wolfgang Macherey, Chris Dyer and Franz Och*
14:10 – 14:35	Forest-based Tree Sequence to String Translation Model *Hui Zhang, Min Zhang, Haizhou Li, Aiti Aw and Chew Lim Tan*
14:35 – 15:00	Active Learning for Multilingual Statistical Machine Translation *Gholamreza Haffari and Anoop Sarkar*

Session 2B (Ballroom 1): Generation and Summarization 1

Session Chair: *Anja Belz*

13:20 – 13:45	DEPEVAL(summ): Dependency-based Evaluation for Automatic Summaries *Karolina Owczarzak*
13:45 – 14:10	Summarizing Definition from Wikipedia *Shiren Ye, Tat-Seng Chua and Jie Lu*
14:10 – 14:35	Automatically Generating Wikipedia Articles: A Structure Aware Approach *Christina Sauper and Regina Barzilay*
14:35 – 15:00	Learning to Tell Tales: A Data-driven Approach to Story Generation *Neil McIntyre and Mirella Lapata*

Session 2C (MR203): Sentiment Analysis & Text Categorization 1

Session Chair: *Katja Markert*

13:20 – 13:45	Recognizing Stances in Online Debates *Swapna Somasundaran and Janyce Wiebe*
13:45 – 14:10	Co-Training for Cross-Lingual Sentiment Classification *Xiaojun Wan*
14:10 – 14:35	A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge *Tao Li, Yi Zhang and Vikas Sindhwani*
14:35 – 15:00	Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis *Jungi Kim, Jin-Ji Li and Jong-Hyeok Lee*

Session 2D (MR205): Language Resources

Session Chair: *Nicoletta Calzolari*

13:20 – 13:45	Compiling a Massive, Multilingual Dictionary via Probabilistic Inference *Mausam, Stephen Soderland, Oren Etzioni, Daniel Weld, Michael Skinner and Jeff Bilmes*
13:45 – 14:10	A Metric-based Framework for Automatic Taxonomy Induction *Hui Yang and Jamie Callan*
14:10 – 14:35	Learning with Annotation Noise *Eyal Beigman and Beata Beigman Klebanov*
14:35 – 15:00	Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both? *Paola Merlo and Lonneke van der Plas*

15:00 – 15:30 Break

Session 3A (Ballroom 2): Machine Translation 2

Session Chair: *Haifeng Wang*

15:30 – 15:55	Robust Machine Translation Evaluation with Entailment Features *Sebastian Pado, Michel Galley, Dan Jurafsky and Christopher D. Manning*
15:55 – 16:20	The Contribution of Linguistic Features to Automatic Machine Translation Evaluation *Enrique Amigó, Jesús Giménez, Julio Gonzalo and Felisa Verdejo*
16:20 – 16:45	A Syntax-Driven Bracketing Model for Phrase-Based Translation *Deyi Xiong, Min Zhang, Aiti Aw and Haizhou Li*
16:45 – 17:10	Topological Ordering of Function Words in Hierarchical Phrase-based Translation *Hendra Setiawan, Min-Yen Kan, Haizhou Li and Philip Resnik*
17:10 – 17:35	Phrase-Based Statistical Machine Translation as a Traveling Salesman Problem *Mikhail Zaslavskiy, Marc Dymetman and Nicola Cancedda*

Session 3B (Ballroom 1): Syntax and Parsing 2

Session Chair: *Dan Klein*

15:30 – 15:55	Concise Integer Linear Programming Formulations for Dependency Parsing *Andre Martins, Noah A. Smith and Eric Xing*
15:55 – 16:20	Non-Projective Dependency Parsing in Expected Linear Time *Joakim Nivre*
16:20 – 16:45	Semi-supervised Learning of Dependency Parsers using Generalized Expectation Criteria *Gregory Druck, Gideon Mann and Andrew McCallum*
16:45 – 17:10	Dependency Grammar Induction via Bitext Projection Constraints *Kuzman Ganchev, Jennifer Gillenwater and Ben Taskar*
17:10 – 17:35	Cross-Domain Dependency Parsing Using a Deep Linguistic Grammar *Yi Zhang and Rui Wang*

Session 3C (MR203): Information Extraction 1

Session Chair: *Eduard Hovy*

15:30 – 15:55	A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment *Fan Yang, Jun Zhao and Kang Liu*
15:55 – 16:20	Reducing Semantic Drift with Bagging and Distributional Similarity *Tara McIntosh and James R. Curran*
16:20 – 16:45	Jointly Identifying Temporal Relations with Markov Logic *Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara and Yuji Matsumoto*
16:45 – 17:10	Profile Based Cross-Document Coreference Using Kernelized Soft Relational Clustering *Jian Huang, Sarah M. Taylor, Jonathan L. Smith, Konstantinos A. Fotiadis and C. Lee Giles*
17:10 – 17:35	Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task *Kristen Parton, Kathleen R. McKeown, Bob Coyne, Mona T. Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gokhan Tur, Wei Xu and Sibel Yaman*

Session 3D (MR209): Semantics 2

Session Chair: *Patrick Pantel*

15:30 – 15:55	Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition *Jong-Hoon Oh, Kiyotaka Uchimoto and Kentaro Torisawa*
15:55 – 16:20	Automatic Set Instance Extraction using the Web *Richard C. Wang and William W. Cohen*
16:20 – 16:45	Extracting Lexical Reference Rules from Wikipedia *Eyal Shnarch, Libby Barak and Ido Dagan*
16:45 – 17:10	Employing Topic Models for Pattern-based Semantic Class Discovery *Huibin Zhang, Mingjie Zhu, Shuming Shi and Ji-Rong Wen*
17:10 – 17:35	Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition *Dipanjan Das and Noah A. Smith*

Software Demonstrations

Session Chairs: *Gary Geunbae Lee and Sabine Schulte im Walde*
Time and Venue: 17:35 – 19:00, Ballroom Foyer, Level 2
Note to demonstrators:

1. Each software demo paper is provided with 1 table, 2 chairs and a poster easel (100cm in width and 200cm in height) good for 1 A0 poster.
2. Two power supply outlets (AC 220V) are available for each table. Please refer to the plug specification.
3. Please set up by 17:20 and take down by 19:15.

D1: MARS: Multilingual Access and Retrieval System with Enhanced Query Translation and Document Retrieval
Lee Lian Hau

D2: A NLG-based Application for Walking Directions
Michael Roth and Anette Frank

D3: A Web-Based Interactive Computer Aided Translation Tool
Philipp Koehn

D4: WISDOM: A Web Information Credibility Analysis System
Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Kentaro Inui, Sadao Kurohashi and Yutaka Kidawara

D5: System for Querying Syntactically Annotated Corpora
Petr Pajas and Jan Stepanek

D6: ProLiV - a Tool for Teaching by Viewing Computational Linguistics
Monica Gavrila and Cristina Vertan

D7: A Tool for Deep Semantic Encoding of Narrative Texts
David K. Elson and Kathleen R. McKeown

D8: Demonstration of Joshua: An Open Source Toolkit for Parsing-based Machine Translation
Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N.G. Thornton, Jonathan Weese and Omar F. Zaidan

D9: WikiBABEL: A Wiki-style Platform for Creation of Parallel Data
A Kumaran and Vikram Dendi

D10: LX-Center: a center of online linguistic services
Antonio Branco, Francisco Costa, Eduardo Ferreira, Pedro Martins, Filipe Nunes, Joao Silva and Sara Silveira

D11: Combining POMDPs trained with User Simulations and Rule-based Dialogue Management in a Spoken Dialogue System
Sebastian Varges, Silvia Quarteroni, Giuseppe Riccardi, Alexei V. Ivanov and Pierluigi Roberti

7

Tuesday, 4 August

Session	Plenary / Oral Session A	Oral Session B	Oral Session C	Oral Session D
Venue	Ballroom 2	Ballroom 1	MR 203	MR 209
8:30-9:45	4A: Statistical and Machine Learning Methods 2	4B: Word Segmentation and POS Tagging	4C: Spoken Language Processing 1	4DI: Short Paper 1 & 2 (Syntax and Parsing, Discourse and Dialogue)
9:45-10:15	AM Coffee/Tea Break (Ballroom Foyer)			
10:15-12:20	5A: Machine Translation 3	5B: Semantics 3	5C: Discourse and Dialogue 1	5D: Student Research Workshop
12:20-14:20	Short Paper Poster/SRW Poster (Lunch) (Ballroom Foyer)			
14:20-16:25	6A/6E: Short Paper 3 & 7 (Machine Translation, Summarization and Generation)	6B/6F: Short Paper 4 & 8 (Semantics, Sentiment Analysis)	6C/6G: Short Paper 5 & 9 (Spoken Language Processing, Question Answering)	6D/6H: Short Paper 6 & 10 (Statistical and Machine Learning Methods 1 & 2)
16:25-16:50	PM Coffee/Tea Break (Ballroom Foyer)			
16:50-18:05	7A: Sentiment Analysis 2	7B: Question Answering	7C: Spoken Language Processing 2	7D: Short paper 11 (Information Extraction)
18:30-19:30	Dinner Cocktail (Level 5, Marina Mandarin Hotel)			
19:30-22:00	Banquet (Level 1, Marina Mandarin Hotel)			

ACL-IJCNLP – Day 2

Session 4A (Ballroom 2): Statistical & Machine Learning Methods 2
Session Chair: *Hal Daume III*

8:30 – 8:55	Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty *Yoshimasa Tsuruoka, Jun'ichi Tsujii and Sophia Ananiadou*
8:55 – 9:20	A global model for joint lemmatization and part-of-speech prediction *Kristina Toutanova and Colin Cherry*
9:20 – 9:45	Distributional Representations for Handling Sparsity in Supervised Sequence-Labeling *Fei Huang and Alexander Yates*

Session 4B (Ballroom 1): Word Segmentation and POS Tagging
Session Chair: *Hwee Tou Ng*

8:30 – 8:55	Minimized Models for Unsupervised Part-of-Speech Tagging *Sujith Ravi and Kevin Knight*
8:55 – 9:20	An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging *Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa and Hitoshi Isahara*
9:20 – 9:45	Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging – A Case Study *Wenbin Jiang, Liang Huang and Qun Liu*

Session 4C (MR203): Spoken Language Processing 1
Session Chair: *Brian Roark*

8:30 – 8:55	Linefeed Insertion into Japanese Spoken Monologue for Captioning *Tomohiro Ohno, Masaki Murata and Shigeki Matsubara*
8:55 – 9:20	Semi-supervised Learning for Automatic Prosodic Event Detection Using Co-training Algorithm *Je Hun Jeon and Yang Liu*
9:20 – 9:45	Summarizing multiple spoken documents: finding evidence from untranscribed audio *Xiaodan Zhu, Gerald Penn and Frank Rudzicz*

Session 4DI (MR209): Short Paper 1 (Syntax and Parsing)
Session Chair: *Joakim Nivre*

8:30 – 8:45	Variational Inference for Grammar Induction with Prior Knowledge *Shay Cohen and Noah A. Smith*
8:45 – 9:00	Bypassed alignment graph for learning coordination in Japanese sentences *Hideharu Okuma, Kazuo Hara, Masashi Shimbo and Yuji Matsumoto*
9:00 – 9:15	An Earley Parsing Algorithm for Range Concatenation Grammars *Laura Kallmeyer, Wolfgang Maier and Yannick Parmentier*

Session 4DII (MR209): Short Paper 2 (Discourse and Dialogue)
Session Chair: *Dragomir R. Radev*

9:15 – 9:30	Using Syntax to Disambiguate Explicit Discourse Connectives in Text *Emily Pitler and Ani Nenkova*
9:30 – 9:45	Hybrid Approach to User Intention Modeling for Dialog Simulation *Sangkeun Jung, Cheongjae Lee, Kyungduk Kim and Gary Geunbae Lee*

9:45 – 10:15	Break

Session 5A (Ballroom 2): Machine Translation 3
Session Chair: *Dekai Wu*

10:15 – 10:40	Improving Tree-to-Tree Translation with Packed Forests *Yang Liu, Yajuan Lü and Qun Liu*
10:40 – 11:05	Fast Consensus Decoding over Translation Forests *John DeNero, David Chiang and Kevin Knight*
11:05 – 11:30	Joint Decoding with Multiple Translation Models *Yang Liu, Haitao Mi, Yang Feng and Qun Liu*
11:30 – 11:55	Collaborative Decoding: Partial Hypothesis Re-ranking Using Translation Consensus between Decoders *Mu Li, Nan Duan, Dongdong Zhang, Chi-Ho Li and Ming Zhou*
11:55 – 12:20	Variational Decoding for Statistical Machine Translation *Zhifei Li, Jason Eisner and Sanjeev Khudanpur*

Session 5B (Ballroom 1): Semantics 3

Session Chair: *Diana McCarthy*

10:15 – 10:40	Unsupervised Learning of Narrative Schemas and their Participants *Nathanael Chambers and Dan Jurafsky*
10:40 – 11:05	Learning a Compositional Semantic Parser using an Existing Syntactic Parser *Ruifang Ge and Raymond Mooney*
11:05 – 11:30	Latent Variable Models of Concept-Attribute Attachment *Joseph Reisinger and Marius Pasca*
11:30 – 11:55	The Chinese Aspect Generation Based on Aspect Selection Functions *Guowen Yang and John Bateman*
11:55 – 12:20	Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation *Kai-min K. Chang, Vladimir L. Cherkassky, Tom M. Mitchell and Marcel Adam Just*

Session 5C (MR203): Discourse and Dialogue 1

Session Chair: *Kathleen R. McKeown*

10:15 – 10:40	Capturing Salience with a Trainable Cache Model for Zero-anaphora Resolution *Ryu Iida, Kentaro Inui and Yuji Matsumoto*
10:40 – 11:05	Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art *Veselin Stoyanov, Nathan Gilbert, Claire Cardie and Ellen Riloff*
11:05 – 11:30	A Novel Discourse Parser Based on Support Vector Machine Classification *David duVerle and Helmut Prendinger*
11:30 – 11:55	Genre distinctions for discourse in the Penn TreeBank *Bonnie Webber*
11:55 – 12:20	Automatic sense prediction for implicit discourse relations in text *Emily Pitler, Annie Louis and Ani Nenkova*

Session 5D (MR209): Student Research Workshop Oral Session

Faculty Advisors: *Grace Ngai and Brian Roark*

Student Chairs: *Davis Dimalen, Jenny Rose Finkel and Blaise Thomson*

10:15 – 10:40	Sense-based Interpretation of Logical Metonymy Using a Statistical Method *Ekaterina Shutova*
10:40 – 11:05	Insights into Non-projectivity in Hindi *Prashanth Mannem and Himani Chaudhry*
11:05 – 11:30	Annotating and Recognising Named Entities in Clinical Notes *Yefeng Wang*
11:30 – 11:55	Paraphrase Recognition Using Machine Learning to Combine Similarity Measures *Prodromos Malakasiotis*
11:55 – 12:20	A System for Semantic Analysis of Chemical Compound Names *Henriette Engelken*
12:20 – 14:20	**Short Paper Poster Session (Lunch)** Session Chairs: *Pushpak Bhattacharyya and Virach Sornlertlamvanich*

Cluster 1: Phonology, Word Segmentation and POS tagging (P1-4)

Homophones and Tonal Patterns in English-Chinese Transliteration
Oi Yee Kwong

Capturing Errors in Written Chinese Words
Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang and Shih-Hung Wu

A Novel Word Segmentation Approach for Written Languages with Word Boundary Markers
Han-Cheol Cho, Do-Gil Lee, Jung-Tae Lee, Pontus Stenetorp, Jun'ichi Tsujii and Hae-Chang Rim

Part of Speech Tagger for Assamese Text
Navanath Saharia, Dhrubajyoti Das, Utpal Sharma and Jugal Kalita

Cluster 2: Syntax and Parsing (P5-9)

Improving data-driven dependency parsing using large-scale LFG grammars
Lilja Øvrelid, Jonas Kuhn and Kathrin Spreyer

12:20 – 14:20 **Short Paper Poster Session (Lunch) (con't)**

Incremental Parsing with Monotonic Adjoining Operation
Yoshihide Kato and Shigeki Matsubara

Bayesian Learning of a Tree Substitution Grammar
Matt Post and Daniel Gildea

A Unified Single Scan Algorithm for Japanese Base Phrase Chunking and Dependency Parsing
Manabu Sassano and Sadao Kurohashi

Comparing the Accuracy of CCG and Penn Treebank Parsers
Stephen Clark and James R. Curran

Cluster 3: Semantics (P10-14)

A Framework for Entailed Relation Recognition
Dan Roth, Mark Sammons and V.G.Vinod Vydiswaran

A Combination of Active Learning and Semi-supervised Learning Starting with Positive and Unlabeled Examples for Word Sense Disambiguation: An Empirical Study on Japanese Web Search Query
Makoto Imamura, Yasuhiro Takayama, Nobuhiro Kaji, Masashi Toyoda and Masaru Kitsuregawa

Detecting Compositionality in Multi-Word Expressions
Ioannis Korkontzelos and Suresh Manandhar

Directional Distributional Similarity for Lexical Expansion
Lili Kotlerman, Ido Dagan, Idan Szpektor and Maayan Zhitomirsky-Geffet

Generalizing over Lexical Features: Selectional Preferences for Semantic Role Classification
Beñat Zapirain, Eneko Agirre and Lluís Màrquez

Cluster 4: Discourse and Dialogue (P15-19)

A Syntactic and Lexical-Based Discourse Segmenter
Milan Tofiloski, Julian Brooke and Maite Taboada

Realistic Grammar Error Simulation using Markov Logic
Sungjin Lee and Gary Geunbae Lee

Discriminative Approach to Predicate-Argument Structure Analysis with Zero-Anaphora Resolution
Kenji Imamura, Kuniko Saito and Tomoko Izumi

Predicting Barge-in Utterance Errors by usingImplicitly-Supervised ASR Accuracy and Barge-in Rate per User
Kazunori Komatani and Alexander I. Rudnicky

Automatic Generation of Information-seeking Questions Using Concept Clusters
Shuguang Li and Suresh Manandhar

12:20 – 14:20 **Short Paper Poster Session (Lunch) (con't)**

Cluster 5: Summarization and Generation (P20-25)

Correlating Human and Automatic Evaluation of a German Surface Realiser
Aoife Cahill

Leveraging Structural Relations for Fluent Compressions at Multiple Compression Rates
Sourish Chaudhuri, Naman K. Gupta, Noah A. Smith and Carolyn P. Rose

Query-Focused Summaries or Query-Biased Summaries?
Rahul Katragadda and Vasudeva Varma

Using Generation for Grammar Analysis and Error Detection
Michael Goodman and Francis Bond

An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation
You Ouyang, Wenjie Li and Qin Lu

Co-Feedback Ranking for Query-Focused Summarization
Furu Wei, Wenjie Li and Yanxiang He

Cluster 6: Machine Translation (P26-32)

Reducing SMT Rule Table with Monolingual Key Phrase
Zhongjun He, Yao Meng, Yajuan Lü, Hao Yu and Qun Liu

A Statistical Machine Translation Model Based on a Synthetic Synchronous Grammar
Hongfei Jiang, Muyun Yang, Tiejun Zhao, Sheng Li and Bo Wang

English-Chinese Bi-Directional OOV Translation based on Web Mining and Supervised Learning
Yuejie Zhang, Yang Wang and Xiangyang Xue

The Backtranslation Score: Automatic MT Evalution at the Sentence Level without Reference Translations
Reinhard Rapp

Sub-Sentence Division for Tree-Based Machine Translation
Hao Xiong, Wenwen Xu, Haitao Mi, Yang Liu and Qun Liu

Asynchronous Binarization for Synchronous Grammars
John DeNero, Adam Pauls and Dan Klein

Hidden Markov Tree Model in Dependency-based Machine Translation
Zdenek Zabokrtsky and Martin Popel

12:20 – 14:20 **Short Paper Poster Session (Lunch) (con't)**

Cluster 7: Sentiment Analysis (P33-40)

Word to Sentence Level Emotion Tagging for Bengali Blogs
Dipankar Das and Sivaji Bandyopadhyay

Extracting Comparative Sentences from Korean Text Documents Using Comparative Lexical Patterns and Machine Learning Techniques
Seon Yang and Youngjoong Ko

Opinion and Generic Question Answering Systems: a Performance Analysis
Alexandra Balahur, Ester Boldrini, Andrés Montoyo and Patricio Martínez-Barco

Automatic Satire Detection: Are You Having a Laugh?
Clint Burfoot and Timothy Baldwin

Hierarchical Multi-Label Text Categorization with Global Margin Maximization
Xipeng Qiu, Wenjun Gao and Xuanjing Huang

Toward finer-grained sentiment identification in product reviews through linguistic and ontological analyses
Hye-Jin Min and Jong C. Park

Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features
Viola Ganter and Michael Strube

Mining User Reviews: from Specification to Summarization
Xinfan Meng and Houfeng Wang

Cluster 8: Information Retrieval (P41-44)

An Ontology-Based Approach for Key Phrase Extraction
Chau Q. Nguyen and Tuoi T. Phan

Query Segmentation Based on Eigenspace Similarity
Chao Zhang, Nan Sun, Xia Hu, Tingzhu Huang and Tat-Seng Chua

Learning Semantic Categories from Clickthrough Logs
Mamoru Komachi, Shimpei Makimoto, Kei Uchiumi and Manabu Sassano

A Rose is a Roos is a Ruusu: Querying Translations for Web Image Search
Janara Christensen, Mausam and Oren Etzioni

12:20 – 14:20 **Short Paper Poster Session (Lunch) (con't)**

Cluster 9: Text Mining and NLP Applications (P45-47)

Extracting Paraphrases of Technical Terms from Noisy Parallel Software Corpora

Xiaoyin Wang, David Lo, Jing Jiang, Lu Zhang and Hong Mei

Mining Association Language Patterns for Negative Life Event Classification

Liang-Chih Yu, Chien-Lung Chan, Chung-Hsien Wu and Chao-Cheng Lin

Automatic Compilation of Travel Information from Automatically Identified Travel Blogs

Hidetsugu Nanba, Haruka Taguma, Takahiro Ozaki, Daisuke Kobayashi, Aya Ishino and Toshiyuki Takezawa

Cluster 10: Language Resources (P48-51)

Play the Language: Play Coreference

Barbora Hladká, Jiří Mírovský and Pavel Schlesinger

Chinese Term Extraction Using Different Types of Relevance

Yuhang Yang, Tiejun Zhao, Qin Lu, Dequan Zheng and Hao Yu

iChi: a bilingual dictionary generating tool

István Varga and Shoichi Yokoyama

CATiB: The Columbia Arabic Treebank

Nizar Habash and Ryan Roth

12:20 – 14:20 **Student Workshop Poster Session (Lunch)**
Faculty Advisors: *Grace Ngai and Brian Roark*
Student Chairs: *Davis Dimalen, Jenny Rose Finkel and Blaise Thomson*

Cluster 11: Student Research Workshop Posters (P52-58)

Sentence diagram generation using dependency parsing
Elijah Mayfield

Accuracy Learning for Chinese Function Tags from Minimal Features
Caixia Yuan

Optimizing Language Model Information Retrieval System with Expectation Maximization Algorithm
Justin Liang-Te Chiu and Jyun-Wei Huang

Data Cleaning for Word Alignment
Tsuyoshi Okita

The Modulation of Cooperation and Emotion in Dialogue: The REC Corpus
Federica Cavicchio

Clustering Technique in Multi-Document Personal Name Disambiguation
Chen Chen and Houfeng Wang

Creating a Gold Standard for Sentence Clustering in Multi-Document Summarization
Johanna Geiss

Session 6A (Ballroom 2): Short Paper 3 (Machine Translation)
Session Chair: *Philipp Koehn*

14:20 – 14:35 A Beam-Search Extraction Algorithm for Comparable Data
Christoph Tillmann

14:35 – 14:50 Optimizing Word Alignment Combination For Phrase Table Training
Yonggang Deng and Bowen Zhou

14:50 – 15:05 Bridging Morpho-Syntactic Gap between Source and Target Sentences for English-Korean Statistical Machine Translation
Gumwon Hong, Seung-Wook Lee and Hae-Chang Rim

15:05 – 15:20 Toward Smaller, Faster, and Better Hierarchical Phrase-based SMT
Mei Yang and Jing Zheng

15:20 – 15:35 Handling phrase reorderings for machine translation
Yizhao Ni, Craig Saunders, Sandor Szedmak and Mahesan Niranjan

Session 6B (Ballroom 1): Short Paper 4 (Semantics)
Session Chair: *Dekang Lin*

14:20 – 14:35 Syntax is from Mars while Semantics from Venus! Insights from Spectral Analysis of Distributional Similarity Networks
Chris Biemann, Monojit Choudhury and Animesh Mukherjee

14:35 – 14:50 Introduction of a new paraphrase generation tool based on Monte-Carlo sampling
Jonathan Chevelu, Thomas Lavergne, Yves Lepage and Thierry Moudenc

14:50 – 15:05 Prediction of Thematic Rank for Structured Semantic Role Labeling
Weiwei Sun, Zhifang Sui and Meng Wang

15:05 – 15:20 Transfer Learning, Feature Selection and Word Sense Disambiguation
Paramveer S. Dhillon and Lyle H. Ungar

Session 6C (MR203): Short Paper 5 (Spoken Language Processing)

Session Chair: *Sanjeev Khudanpur*

14:20 – 14:35	From Extractive to Abstractive Meeting Summaries: Can It Be Done by Sentence Compression? *Fei Liu and Yang Liu*
14:35 – 14:50	Automatic Story Segmentation using a Bayesian Decision Framework for Statistical Models of Lexical Chain Features *Wai-Kit Lo, Wenying Xiong and Helen Meng*
14:50 – 15:05	Investigating Pitch Accent Recognition in Non-native Speech *Gina-Anne Levow*
15:05 – 15:20	A Stochastic Finite-State Morphological Parser for Turkish *Haşim Sak, Tunga Güngör and Murat Saraçlar*
15:20 – 15:35	Parsing Speech Repair without Specialized Grammar Symbols *Tim Miller, Luan Nguyen and William Schuler*

Session 6D (MR209): Short Paper 6 (Statistical and Machine Learning Methods 1)

Session Chair: *Yuji Matsumoto*

14:20 – 14:35	Efficient Inference of CRFs for Large-Scale Natural Language Data *Minwoo Jeong, Chin-Yew Lin and Gary Geunbae Lee*
14:35 – 14:50	Iterative Scaling and Coordinate Descent Methods for Maximum Entropy *Fang-Lan Huang, Cho-Jui Hsieh, Kai-Wei Chang and Chih-Jen Lin*
14:50 – 15:05	Automatic Cost Estimation for Tree Edit Distance Using Particle Swarm Optimization *Yashar Mehdad*
15:05 – 15:20	Markov Random Topic Fields *Hal Daume III*

Session 6E (Ballroom 2): Short Paper 7 (Summarization and Generation)

Session Chair: *Diana Inkpen*

15:40 – 15:55	Multi-Document Summarization using Sentence-based Topic Models *Dingding Wang, Shenghuo Zhu, Tao Li and Yihong Gong*
15:55 – 16:10	Validating the web-based evaluation of NLG systems *Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Sara Dalzel-Job, Johanna Moore and Jon Oberlander*
16:10 – 16:25	Extending a Surface Realizer to Generate Coherent Discourse *Eva Banik*

Session 6F (Ballroom 1): Short Paper 8 (Sentiment Analysis)

Session Chair: *Bo Pang*

15:25 – 15:40	The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language *Rada Mihalcea and Carlo Strapparava*
15:40 – 15:55	Generalizing Dependency Features for Opinion Mining *Mahesh Joshi and Carolyn Penstein-Ros*
15:55 – 16:10	Graph Ranking for Sentiment Transfer *Qiong Wu, Songbo Tan and Xueqi Cheng*
16:10 – 16:25	The Contribution of Stylistic Information to Content-based Mobile Spam Filtering *Dae-Neung Sohn, Jung-Tae Lee and Hae-Chang Rim*

Session 6G (MR203): Short Paper 9 (Question Answering)

Session Chair: *Tomek Strzalkowski*

15:40 – 15:55	Learning foci for Question Answering over Topic Maps *Alexander Mikhailian, Tiphaine Dalmas and Rani Pinchuk*
15:55 – 16:10	Do Automatic Annotation Techniques Have Any Impact on Supervised Complex Question Answering? *Yllias Chali, Sadid Hasan and Shafiq Joty*
16:10 – 16:25	Where's the Verb? Correcting Machine Translation During Question Answering *Wei-Yun Ma and Kathleen R. McKeown*

Session 6H (MR209): Short Paper 10 (Statistical and Machine Learning Methods 2)

Session Chair: *Kenneth Church*

15:25 – 15:40	A Note on the Implementation of Hierarchical Dirichlet Processes *Phil Blunsom, Trevor Cohn, Sharon Goldwater and Mark Johnson*
15:40 – 15:55	A Succinct N-gram Language Model *Taro Watanabe, Hajime Tsukada and Hideki Isozaki*
15:55 – 16:10	Modeling Morphologically Rich Languages Using Split Words and Unstructured Dependencies *Deniz Yuret and Ergun Bicici*
16:10 – 16:25	Improved Smoothing for N-gram Language Models Based on Ordinary Counts *Robert C. Moore and Chris Quirk*
16:25 – 16:50	Break

Session 7A (Ballroom 2): Sentiment Analysis & Text Categorization 2

Session Chair: *Bing Liu*

16:50 – 17:15	A Framework of Feature Selection Methods for Text Categorization *Shoushan Li, Rui Xia, Chengqing Zong and Chu-Ren Huang*
17:15 – 17:40	Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification *Sajib Dasgupta and Vincent Ng*
17:40 – 18:05	Modeling Latent Biographic Attributes in Conversational Genres *Nikesh Garera and David Yarowsky*

Session 7B (Ballroom 1): Question Answering

Session Chair: *Chin-Yew Lin*

16:50 – 17:15	A Graph-based Semi-Supervised Learning for Question-Answering *Asli Celikyilmaz, Marcus Thint and Zhiheng Huang*
17:15 – 17:40	Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding *Delphine Bernhard and Iryna Gurevych*
17:40 – 18:05	Answering Opinion Questions with Random Walks on Graphs *Fangtao Li, Yang Tang, Minlie Huang and Xiaoyan Zhu*

Session 7C (MR203): Spoken Language Processing 2

Session Chair: *Yang Liu*

16:50 – 17:15	What lies beneath: Semantic and syntactic analysis of manually reconstructed spontaneous speech *Erin Fitzgerald, Frederick Jelinek and Robert Frank*
17:15 – 17:40	Discriminative Lexicon Adaptation for Improved Character Accuracy - A New Direction in Chinese Language Modeling *Yi-Cheng Pan, Lin-Shan Lee and Sadaoki Furui*
17:40 – 18:05	Improving Automatic Speech Recognition for Lectures through Transformation-based Rules Learned from Minimal Data *Cosmin Munteanu, Gerald Penn and Xiaodan Zhu*

Session 7D (MR209): Short paper 11 (Information Extraction)

Session Chair: *Raymond Mooney*

16:50 – 17:05	Updating a Name Tagger Using Contemporary Unlabeled Data *Cristina Mota and Ralph Grishman*
17:05 – 17:20	Arabic Cross-Document Coreference Resolution *Asad Sayeed, Tamer Elsayed, Nikesh Garera, David Alexander, Tan Xu, Doug Oard, David Yarowsky and Christine Piatko*
17:20 – 17:35	The Impact of Query Refinement in the Web People Search Task *Javier Artiles, Julio Gonzalo and Enrique Amigó*
17:35 – 17:50	Composite Kernels For Relation Extraction *Frank Reichartz, Hannes Korte and Gerhard Paass*
17:50 – 18:05	Predicting Unknown Time Arguments based on Cross-Event Propagation *Prashant Gupta and Heng Ji*

8

Wednesday, 5 August

Session	Plenary / Oral Session A	Oral Session B	Oral Session C	Oral Session D
Venue	Ballroom 2	Ballroom 1	MR 203	MR 209
8:30-9:30	Invited Talk 2			
9:30-9:55	AM Coffee/Tea Break (Ballroom Foyer)			
9:55-11:35	8A: Machine Translation 4	8B: Generation and Summarization 2	8C: Text Mining and NLP Applications	8D: Disourse and Dialogue 2
11:35-12:30	Lunch			
12:30-14:00	ACL Business Meeting			
14:00-14:25	PM Coffee/Tea Break (Ballroom Foyer)			
14:25-16:30	9A: Machine Translation 5	9B: Syntax and Parsing 3	9C: Information Extraction 2	9D: Information Retrieval
16:40-18:15	Lifetime Achievement Award and Presentation			
18:15-18:45	Closing Session: Best Paper Awards and Future Conferences			

Invited Talk

Wednesday, 5 August
8:30 – 9:30, Ballroom 2

Bonnie Webber, University of Edinburgh, UK

Discourse – Early Problems, Current Successes, Future Challenges

I will look back through nearly forty years of computational research on discourse, noting some problems (such as context-dependence and inference) that were identified early on as a hindrance to further progress, some admirable successes that we have achieved so far in the development of algorithms and resources, and some challenges that we may want to (or that we may have to!) take up in the future, with particular attention to problems of data annotation and genre dependence.

Bonnie Webber was a researcher at Bolt Beranek and Newman while working on the PhD she received from Harvard University in 1978. She then taught in the Department of Computer and Information Science at the University of Pennsylvania for 20 years before joining the School of Informatics at the University of Edinburgh. Known for research on discourse and on question answering, she is a Past President of the Association for Computational Linguistics, co-developer (with Aravind Joshi, Rashmi Prasad, Alan Lee and Eleni Miltsakaki) of the Penn Discourse TreeBank, and co-editor (with Annie Zaenen and Martha Palmer) of the journal, *Linguistic Issues in Language Technology*.

ACL-IJCNLP – Day 3

8:30 – 9:30	Invited Talk: Discourse – Early Problems, Current Successes, Future Challenges (Ballroom 2) *Bonnie Webber*
9:30 – 10:00	Break

Session 8A (Ballroom 2): Machine Translation 4

Session Chair: *Dan Gildea*

9:55 – 10:20	Quadratic-Time Dependency Parsing for Machine Translation *Michel Galley and Christopher D. Manning*
10:20 – 10:45	A Gibbs Sampler for Phrasal Synchronous Grammar Induction *Phil Blunsom, Trevor Cohn, Chris Dyer and Miles Osborne*
10:45 – 11:10	Source-Language Entailment Modeling for Translating Unknown Terms *Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman and Idan Szpektor*
11:10 – 11:35	Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT *Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh and Pushpak Bhattacharyya*

Session 8B (Ballroom 1): Generation and Summarization 2

Session Chair: *Regina Barzilay*

9:55 – 10:20	Dependency Based Chinese Sentence Realization *Wei He, Haifeng Wang, Yuqing Guo and Ting Liu*
10:20 – 10:45	Incorporating Information Status into Generation Ranking *Aoife Cahill and Arndt Riester*
10:45 – 11:10	A Syntax-Free Approach to Japanese Sentence Compression *Tsutomu Hirao, Jun Suzuki and Hideki Isozaki*
11:10 – 11:35	Application-driven Statistical Paraphrase Generation *Shiqi Zhao, Xiang Lan, Ting Liu and Sheng Li*

Session 8C (MR203): Text Mining and NLP applications
Session Chair: *Sophia Ananioudou*

9:55 – 10:20	Semi-Supervised Cause Identification from Aviation Safety Reports *Isaac Persing and Vincent Ng*
10:20 – 10:45	SMS based Interface for FAQ Retrieval *Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkatesan T. Chakaravarthy and L. Venkata Subramaniam*
10:45 – 11:10	Semantic Tagging of Web Search Queries *Mehdi Manshadi and Xiao Li*
11:10 – 11:35	Mining Bilingual Data from the Web with Adaptively Learnt Patterns *Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu and Qingsheng Zhu*

Session 8D (MR209): Discourse and Dialogue 2
Session Chair: *Gary Geunbae Lee*

9:55 – 10:20	Comparing Objective and Subjective Measures of Usability in a Human-Robot Dialogue System *Mary Ellen Foster, Manuel Giuliani and Alois Knoll*
10:20 – 10:45	Setting Up User Action Probabilities in User Simulations for Dialog System Development *Hua Ai and Diane Litman*
10:45 – 11:10	Dialogue Segmentation with Large Numbers of Volunteer Internet Annotators *T. Daniel Midgley*
11:10 – 11:35	Robust Approach to Abbreviating Terms: A Discriminative Latent Variable Model with Global Information *Xu Sun, Naoaki Okazaki and Jun'ichi Tsujii*

11:35 – 12:30	Lunch
12:30 – 14:00	ACL Business Meeting (Ballroom 2)
14:00 – 14:25	Break

Session 9A (Ballroom 2): Machine Translation 5
Session Chair: *Kevin Knight*

14:25 – 14:50	A non-contiguous Tree Sequence Alignment-based Model for Statistical Machine Translation *Jun Sun, Min Zhang and Chew Lim Tan*
14:50 – 15:15	Better Word Alignments with Supervised ITG Models *Aria Haghighi, John Blitzer, John DeNero and Dan Klein*
15:15 – 15:40	Confidence Measure for Word Alignment *Fei Huang*
15:40 – 16:05	A Comparative Study of Hypothesis Alignment and its Improvement for Machine Translation System Combination *Boxing Chen, Min Zhang, Haizhou Li and Aiti Aw*
16:05 – 16:30	Incremental HMM Alignment for MT System Combination *Chi-Ho Li, Xiaodong He, Yupeng Liu and Ning Xi*

Session 9B (Ballroom 1): Syntax and Parsing 3
Session Chair: *James R. Curran*

14:25 – 14:50	K-Best A* Parsing *Adam Pauls and Dan Klein*
14:50 – 15:15	Coordinate Structure Analysis with Global Structural Constraints and Alignment-Based Local Features *Kazuo Hara, Masashi Shimbo, Hideharu Okuma and Yuji Matsumoto*
15:15 – 15:40	Learning Context-Dependent Mappings from Sentences to Logical Form *Luke Zettlemoyer and Michael Collins*
15:40 – 16:05	An Optimal-Time Binarization Algorithm for Linear Context-Free Rewriting Systems with Fan-Out Two *Carlos Gómez-Rodríguez and Giorgio Satta*
16:05 – 16:30	A Polynomial-Time Parsing Algorithm for TT-MCTAG *Laura Kallmeyer and Giorgio Satta*

Session 9C (MR203): Information Extraction 2

Session Chair: *Jian Su*

14:25 – 14:50	Distant supervision for relation extraction without labeled data *Mike Mintz, Steven Bills, Rion Snow and Dan Jurafsky*
14:50 – 15:15	Multi-Task Transfer Learning for Weakly-Supervised Relation Extraction *Jing Jiang*
15:15 – 15:40	Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web *Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang and Mitsuru Ishizuka*
15:40 – 16:05	Phrase Clustering for Discriminative Learning *Dekang Lin and Xiaoyun Wu*
16:05 – 16:30	Semi-Supervised Active Learning for Sequence Labeling *Katrin Tomanek and Udo Hahn*

Session 9D (MR209): Information Retrieval

Session Chair: *Kam-Fai Wong*

14:25 – 14:50	Word or Phrase? Learning Which Unit to Stress for Information Retrieval *Young-In Song, Jung-Tae Lee and Hae-Chang Rim*
14:50 – 15:15	A Generative Blog Post Retrieval Model that Uses Query Expansion based on External Collections *Wouter Weerkamp, Krisztian Balog and Maarten de Rijke*
15:15 – 15:40	Language Identification of Search Engine Queries *Hakan Ceylan and Yookyung Kim*
15:40 – 16:05	Exploiting Bilingual Information to Improve Web Search *Wei Gao, John Blitzer, Ming Zhou and Kam-Fai Wong*

16:40 – 18:15	Lifetime Achievement Award and Presentation (Ballrooom 2)
18:15 – 18:45	Closing Session: Best Paper Awards, Future Conferences (Ballroom 2)

9

Thursday, 6 August

Session	EMNLP Oral A	EMNLP Oral B	EMNLP Oral C	EMNLP Oral D	Workshops
Venue	Theater	MR208	MR209	MR203	Various MR, Level 3
8:30-10:00	Opening Remarks / Invited Talk				Morning Session 1
10:00-10:30	AM Coffee/Tea Break (Ballroom Foyer)				
10:30-12:10	1A: Semantic Parsing	1B: Machine Translation I	1C: Machine Learning and Statistical Models I	Information Extraction	Morning Session 2
12:10-13:50	Lunch				
13:50-15:30	2A: Subjectivity and Sentiment I	2B: Machine Translation II	2C: Natural Language Processing for Web 2.0	2D: Language Resources and Evaluation	Afternoon Session 1
15:30-16:00	PM Coffee/Tea Break (Ballroom Foyer)				
16:00-18:00	3A: Discourse and Dialogue	3B: Machine Translation III	3C: Summarization and Generation	3D: Lexical Semantics I	Afternoon Session 2
18:00-20:00	EMNLP Poster Session and Reception				

Invited Talk

Thursday, 6 August
9:00-10:00, Theatre

Richard Sproat, Oregon Health & Science University

Symbols, Meaning and Statistics

Of all artifacts left behind by ancient cultures, few are as evocative as symbols. When one looks at an inscription on stone or clay, it is natural to ask: What does it mean? Was this a form of writing, or some sort of non-linguistic system? If it was writing, can we hope to decipher it?

In this talk, I examine these and related questions, and the possible role of statistical methods in answering them. I start with some highlights from the history of successful decipherment. I review work on computational approaches to decipherment, and assess how useful this is likely to be. One area where it is clearly useful, in a sense, is in generating pseudodecipherments, and I present one such case as a reductio ad absurdum of attempts to decipher artifacts like the Phaistos disk. And I discuss a topic that made its rounds of the popular science press earlier this year: Namely, the claimed "entropic evidence" that the 4000-year-old Indus Valley symbol system constituted a script. I present simple counterevidence to the usefulness of the proposed measure to support any such claim; and I review the large amount of archaeological and comparative cultural evidence against the script hypothesis, which must in any case be taken into account in any complete discussion of this subject.

(Portions of this talk are based on joint work with Steve Farmer and Michael Witzel, and on a tutorial at NAACL 2009 co-presented with Kevin Knight.)

Richard Sproat received his Ph.D. in Linguistics from the Massachusetts Institute of Technology in 1985. He worked at AT&T (Bell) Labs, before moving to the University of Illinois (2003) and thence to the Oregon Health & Science University (2009). Sproat has worked in a number areas relating to language and computational linguistics, including morphology, computational morphology, articulatory and acoustic phonetics, text processing, text-to-speech synthesis, speech recognition and text-to-scene conversion. One of his long-term interests is writing systems and how they encode language. Since 2004 he has been involved in an often heated debate about the nature of the cryptic symbols left behind by the 4000-year-old Indus Valley civilization.

EMNLP – Day 1

8:45 – 9:00	Opening Remarks
9:00 – 10:00	Invited Talk
10:00 – 10:30	Coffee Break

Session 1A (Theatre): Semantic Parsing

10:30 – 10:55	Unsupervised Semantic Parsing *Hoifung Poon and Pedro Domingos*
10:55 – 11:20	Graph Alignment for Semi-Supervised Semantic Role Labeling *Hagen Fürstenau and Mirella Lapata*
11:20 – 11:45	Semi-supervised Semantic Role Labeling Using the Latent Words Language Model *Koen Deschacht and Marie-Francine Moens*
11:45 – 12:10	Semantic Dependency Parsing of NomBank and PropBank: An Efficient Integrated Approach via a Large-scale Feature Selection *Hai Zhao, Wenliang Chen and Chunyu Kit*

Session 1B (MR208): Machine Translation I

10:30 – 10:55	First- and Second-Order Expectation Semirings with Applications to Minimum-Risk Training on Translation Forests *Zhifei Li and Jason Eisner*
10:55 – 11:20	Feasibility of Human-in-the-loop Minimum Error Rate Training *Omar F. Zaidan and Chris Callison-Burch*
11:20 – 11:45	Cube Pruning as Heuristic Search *Mark Hopkins and Greg Langmead*
11:45 – 12:10	Effective Use of Linguistic and Contextual Information for Statistical Machine Translation *Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas and Ralph Weischedel*

Session 1C (MR209): Machine Learning and Statistical Models I

10:30 – 10:55	Active Learning by Labeling Features *Gregory Druck, Burr Settles and Andrew McCallum*
10:55 – 11:20	Efficient kernels for sentence pair classification *Fabio Massimo Zanzotto and Lorenzo Dell'Arciprete*
11:20 – 11:45	Graphical Models over Multiple Strings *Markus Dreyer and Jason Eisner*
11:45 – 12:10	Reverse Engineering of Tree Kernel Feature Spaces *Daniele Pighin and Alessandro Moschitti*

Session 1D (MR203): Information Extraction

10:30 – 10:55	A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora *Makoto Miwa, Rune Sætre, Yusuke Miyao and Jun'ichi Tsujii*
10:55 – 11:20	Generalized Expectation Criteria for Bootstrapping Extractors using Record-Text Alignment *Kedar Bellare and Andrew McCallum*
11:20 – 11:45	Nested Named Entity Recognition *Jenny Rose Finkel and Christopher D. Manning*
11:45 – 12:10	A Unified Model of Phrasal and Sentential Evidence for Information Extraction *Siddharth Patwardhan and Ellen Riloff*

12:10 – 13:50	Lunch

Session 2A (Theatre): Subjectivity and Sentiment I

13:50 – 14:15	Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm *Jingjing Liu and Stephanie Seneff*
14:15 – 14:40	Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification *Swapna Somasundaran, Galileo Namata, Janyce Wiebe and Lise Getoor*
14:40 – 15:05	Sentiment Analysis of Conditional Sentences *Ramanathan Narayanan, Bing Liu and Alok Choudhary*
15:05 – 15:30	Subjectivity Word Sense Disambiguation *Cem Akkaya, Janyce Wiebe and Rada Mihalcea*

Session 2B (MR208): Machine Translation II

13:50 – 14:15	Non-Projective Parsing for Statistical Machine Translation *Xavier Carreras and Michael Collins*
14:15 – 14:40	Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models *Arne Mauser, Saša Hasan and Hermann Ney*
14:40 – 15:05	Feature-Rich Translation by Quasi-Synchronous Lattice Parsing *Kevin Gimpel and Noah A. Smith*
15:05 – 15:30	Improved Word Alignment with Statistics and Linguistic Heuristics *Ulf Hermjakob*

Session 2C (MR209): Natural Language Processing for Web 2.0

13:50 – 14:15	Entity Extraction via Ensemble Semantics *Marco Pennacchiotti and Patrick Pantel*
14:15 – 14:40	Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora *Daniel Ramage, David Hall, Ramesh Nallapati and Christopher D. Manning*
14:40 – 15:05	Clustering to Find Exemplar Terms for Keyphrase Extraction *Zhiyuan Liu, Peng Li, Yabin Zheng and Maosong Sun*
15:05 – 15:30	Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns *Dmitry Davidov and Ari Rappoport*

Session 2D (MR203): Language Resources and Evaluation

13:50 – 14:15	Wikipedia as Frame Information Repository *Sara Tonelli and Claudio Giuliano*
14:15 – 14:40	Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk *Chris Callison-Burch*
14:40 – 15:05	How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. *Jason Baldridge and Alexis Palmer*
15:05 – 15:30	Automatically Evaluating Content Selection in Summarization without Human Models *Annie Louis and Ani Nenkova*

15:30 – 16:00	Coffee Break

Session 3A (Theatre): Discourse and Dialogue

16:00 – 16:25 Classifier Combination for Contextual Idiom Detection Without Labelled Data
Linlin Li and Caroline Sporleder

16:25 – 16:50 Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing
Brian Roark, Asaf Bachrach, Carlos Cardenas and Christophe Pallier

16:50 – 17:15 It's Not You, it's Me: Detecting Flirting and its Misperception in Speed-Dates
Rajesh Ranganath, Dan Jurafsky and Dan McFarland

17:15 – 17:40 Recognizing Implicit Discourse Relations in the Penn Discourse Treebank
Ziheng Lin, Min-Yen Kan and Hwee Tou Ng

Session 3B (MR208): Machine Translation III

16:00 – 16:25 A Bayesian Model of Syntax-Directed Tree to String Grammar Induction
Trevor Cohn and Phil Blunsom

16:25 – 16:50 Better Synchronous Binarization for Machine Translation
Tong Xiao, Mu Li, Dongdong Zhang, Jingbo Zhu and Ming Zhou

16:50 – 17:15 Accuracy-Based Scoring for DOT: Towards Direct Error Minimization for Data-Oriented Translation
Daniel Galron, Sergio Penkale, Andy Way and I. Dan Melamed

17:15 – 17:40 Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases
Yuval Marton, Chris Callison-Burch and Philip Resnik

Session 3C (MR209): Summarization and Generation

16:00 – 16:25	A Comparison of Model Free versus Model Intensive Approaches to Sentence Compression *Tadashi Nomoto*
16:25 – 16:50	Natural Language Generation with Tree Conditional Random Fields *Wei Lu, Hwee Tou Ng and Wee Sun Lee*
16:50 – 17:15	Perceptron Reranking for CCG Realization *Michael White and Rajakrishnan Rajkumar*
17:15 – 17:40	Multi-Document Summarisation Using Generic Relation Extraction *Ben Hachey*

Session 3D (MR203): Lexical Semantics I

16:00 – 16:25	Language Models Based on Semantic Composition *Jeff Mitchell and Mirella Lapata*
16:25 – 16:50	Graded Word Sense Assignment *Katrin Erk and Diana McCarthy*
16:50 – 17:15	Joint Learning of Preposition Senses and Semantic Roles of Prepositional Phrases *Daniel Dahlmeier, Hwee Tou Ng and Tanja Schultz*
17:15 – 17:40	Projecting Parameters for Multilingual Word Sense Disambiguation *Mitesh M. Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya*

18:00 – 20:00	Poster Session and Reception

EMNLP Posters

Gazpacho and summer rash: lexical relationships from temporal patterns of web search queries
Enrique Alfonseca, Massimiliano Ciaramita and Keith Hall

A Compact Forest for Scalable Inference over Entailment and Paraphrase Rules
Roy Bar-Haim, Jonathan Berant and Ido Dagan

Discriminative Substring Decoding for Transliteration
Colin Cherry and Hisami Suzuki

Re-Ranking Models Based-on Small Training Data for Spoken Language Understanding
Marco Dinarelli, Alessandro Moschitti and Giuseppe Riccardi

Empirical Exploitation of Click Data for Task Specific Ranking
Anlei Dong, Yi Chang, Shihao Ji, Ciya Liao, Xin Li and Zhaohui Zheng

The Feature Subspace Method for SMT System Combination
Nan Duan, Mu Li, Tong Xiao and Ming Zhou

Lattice-based System Combination for Statistical Machine Translation
Yang Feng, Yang Liu, Haitao Mi, Qun Liu and Yajuan Lü

A Joint Language Model With Fine-grain Syntactic Tags
Denis Filimonov and Mary Harper

Bidirectional Phrase-based Statistical Machine Translation
Andrew Finch and Eiichiro Sumita

Real-time decision detection in multi-party dialogue
Matthew Frampton, Jia Huang, Trung Bui and Stanley Peters

On the Role of Lexical Features in Sequence Labeling
Yoav Goldberg and Michael Elhadad

Simple Coreference Resolution with Rich Syntactic and Semantic Features
Aria Haghighi and Dan Klein

Descriptive and Empirical Approaches to Capturing Underlying Dependencies among Parsing Errors
Tadayoshi Hara, Yusuke Miyao and Jun'ichi Tsujii

Large-Scale Verb Entailment Acquisition from the Web
Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Stijn De Saeger, Masaki Murata and Jun'ichi Kazama

A Syntactified Direct Translation Model with Linear-time Decoding
Hany Hassan, Khalil Sima'an and Andy Way

Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge
Samer Hassan and Rada Mihalcea

Joint Optimization for Machine Translation System Combination
Xiaodong He and Kristina Toutanova

Fully Lexicalising CCGbank with Hat Categories
Matthew Honnibal and James R. Curran

Bilingually-Constrained (Monolingual) Shift-Reduce Parsing
Liang Huang, Wenbin Jiang and Qun Liu

Accurate Semantic Class Classifier for Coreference Resolution
Zhiheng Huang, Guangping Zeng, Weiqun Xu and Asli Celikyilmaz

Real-Word Spelling Correction using Google Web 1T 3-grams
Aminul Islam and Diana Inkpen

Semi-supervised Speech Act Recognition in Emails and Forums
Minwoo Jeong, Chin-Yew Lin and Gary Geunbae Lee

Using Morphological and Syntactic Structures for Chinese Opinion Analysis
Lun-Wei Ku, Ting-Hao Huang and Hsin-Hsi Chen

Finding Short Definitions of Terms on Web Pages
Gerasimos Lampouras and Ion Androutsopoulos

Improving Nominal SRL in Chinese Language with Verbal SRL Information and Automatic Predicate Recognition
Junhui Li, Guodong Zhou, Hai Zhao, Qiaoming Zhu and Peide Qian

On the Use of Virtual Evidence in Conditional Random Fields
Xiao Li

Refining Grammars for Parsing with Hierarchical Semantic Knowledge
Xiaojun Lin, Yang Fan, Meng Zhang, Xihong Wu and Huisheng Chi

Bayesian Learning of Phrasal Tree-to-String Templates
Ding Liu and Daniel Gildea

Human-competitive tagging using automatic keyphrase extraction
Olena Medelyan, Eibe Frank and Ian H. Witten

Supervised Learning of a Probabilistic Lexicon of Verb Semantic Classes
Yusuke Miyao and Jun'ichi Tsujii

A Study on the Semantic Relatedness of Query and Document Terms in Information Retrieval
Christof Müller and Iryna Gurevych

Predicting Subjectivity in Multimodal Conversations
Gabriel Murray and Giuseppe Carenini

Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages
Preslav Nakov and Hwee Tou Ng

What's in a name? In some languages, grammatical gender
Vivi Nastase and Marius Popescu

Dependency and Sequential Structures for Relation Extraction Convolution Kernels on Constituent
Truc-Vien T. Nguyen, Alessandro Moschitti and Giuseppe Riccardi

Automatic Acquisition of the *Argument-Predicate* Relations from a Frame-Annotated Corpus
Ekaterina Ovchinnikova, Theodore Alexandrov and Tonio Wandmacher

Detecting Speculations and their Scopes in Scientific Text
Arzucan Özgür and Dragomir R. Radev

Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models
Michael Paul and Roxana Girju

Consensus Training for Consensus Decoding in Machine Translation
Adam Pauls, John DeNero and Dan Klein

Using Word-Sense Disambiguation Methods to Classify Web Queries by Intent
Emily Pitler and Ken Church

Semi-Supervised Learning for Semantic Relation Classification using Stratified Sampling Strategy
Longhua Qian, Guodong Zhou, Fang Kong and Qiaoming Zhu

Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis
Changqin Quan and Fuji Ren

A Probabilistic Model for Associative Anaphora Resolution
Ryohei Sasano and Sadao Kurohashi

Quantifier Scope Disambiguation Using Extracted Pragmatic Knowledge: Preliminary Results
Prakash Srinivasan and Alexander Yates

Chinese Semantic Role Labeling with Shallow Parsing
Weiwei Sun, Zhifang Sui, Meng Wang and Xin Wang

Discovery of Term Variation in Japanese Web Search Queries
Hisami Suzuki, Xiao Li and Jianfeng Gao

Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics
Simone Teufel, Advaith Siddharthan and Colin Batchelor

Character-level Analysis of Semi-Structured Documents for Set Expansion
Richard C. Wang and William W. Cohen

Classifying Relations for Biomedical Named Entity Disambiguation
Xinglong Wang, Jun'ichi Tsujii and Sophia Ananiadou

Domain adaptive bootstrapping for named entity recognition
Dan Wu, Wee Sun Lee, Nan Ye and Hai Leong Chieu

Phrase Dependency Parsing for Opinion Mining
Yuanbin Wu, Qi Zhang, Xuangjing Huang and Lide Wu

Polynomial to Linear: Efficient Classification with Conjunctive Features
Naoki Yoshinaga and Masaru Kitsuregawa

K-Best Combination of Syntactic Parsers
Hui Zhang, Min Zhang, Chew Lim Tan and Haizhou Li

Chinese Novelty Mining
Yi Zhang and Flora S. Tsai

Latent Document Re-Ranking
Dong Zhou and Vincent Wade

Workshop Listing

6-7 August – 2 day workshops

WS6: The Third Linguistic Annotation Workshop (LAW III)
Venue: MR 303
Chairs: Nancy Ide, Adam Meyers, Manfred Stede, Chu-Ren Huang

WS10: The 7th Workshop on Asian Language Resources (ALR7)
Venue: MR302
Chairs: Hammam Riza, Virach Sornlertlamvanich

6 August – 1 day workshops

WS1: Applied Textual Inference (TextInfer)
Venue: MR314
Chairs: Chris Callison-Burch, Ido Dagan, Christopher Manning, Marco Pennacchiotti, Fabio Massimo Zanzotto

WS2: Grammar Engineering across Frameworks (GEAF)
Venue: MR307
Chairs: Tracy Holloway King, Marianne Santaholma

WS3: Knowledge and Reasoning for Answering Questions (KRAQ)
Venue: MR306
Chairs: Marie-Francine Moens, Patrick Saint-Dizier
(Note: KRAQ is a half day workshop, in the afternoon only)

WS4: Language Generation and Summarisation (UCNLG+Sum)
Venue: MR305
Chairs: Anja Belz, Roger Evans, Sebastian Varges

WS5: Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE)
Venue: MR304
Chairs: Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov, Su Nam Kim

WS7: 2nd Workshop on Building and Using Comparable Corpora: from parallel to non-parallel corpora (BUCC)
Venue: MR301
Chairs: Pascale Fung, Pierre Zweigenbaum, Reinhard Rapp

WS1: Applied Textual Inference (TextInfer)

8:45 – 9:00	Opening Remarks

Session 1: Foundational Aspects and Linguistic Analysis of Textual Entailment

9:00 – 9:30	Multi-word expressions in textual inference: Much ado about nothing? *Marie-Catherine de Marneffe, Sebastian Pado and Christopher D. Manning*
9:30 – 10:00	A Proposal on Evaluation Measures for RTE *Richard Bergmair*
10:00 – 10:30	Coffee Break
10:30 – 11:00	Sub-sentencial Paraphrasing by Contextual Pivot Translation *Aurélien Max*
11:00 – 12:00	Invited Talks
12:00 – 13:50	Lunch

Session 2: Learning Textual Entailment Rules and Building Corpora

13:50 – 14:20	Augmenting WordNet-based Inference with Argument Mapping *Idan Szpektor and Ido Dagan*
14:20 – 14:50	Optimizing Textual Entailment Recognition Using Particle Swarm Optimization *Yashar Mehdad and Bernardo Magnini*
14:50 – 15:10	Ranking Paraphrases in Context *Stefan Thater, Georgiana Dinu and Manfred Pinkal*
15:10 – 15:30	Building an Annotated Textual Inference Corpus for Motion and Space *Kirk Roberts*
15:30 – 16:00	Coffee Break

Session 3: Machine Learning Models and Application of Textual Inference

16:00 – 16:30	Using Hypernymy Acquisition to Tackle (Part of) Textual Entailment *Elena Akhmatova and Mark Dras*
16:30 – 17:00	Automating Model Building in c-rater *Jana Sukkarieh and Svetlana Stoyanchev*
17:00 – 17:20	Presupposed Content and Entailments in Natural Language Inference *David Clausen and Christopher D. Manning*
17:20 – 18:00	Final Panel and Discussion

WS2: Grammar Engineering across Frameworks (GEAF)

Session 1

8:30 – 9:00	Exploration of the LTAG-Spinal Formalism and Treebank for Semantic Role Labeling *Yudong Liu and Anoop Sarkar*
9:00 – 9:30	Developing German Semantics on the basis of Parallel LFG Grammars *Sina Zarrieß*
9:30 – 10:00	Mining of Parsed Data to Derive Deverbal Argument Structure *Olga Gurevich and Scott A. Waterman*
10:00 – 10:30	Break
10:30 – 12:10	Demo Session (with quick fire presentations)

Parallel Grammar Engineering for Slavic Languages
Tania Avgustinova and Yi Zhang

HaG — An HPSG of Hausa
Berthold Crysmann

EGAD: Erroneous Generation Analysis and Detection
Michael Goodman and Francis Bond

Deverbal Nouns in a Semantic Search Application
Olga Gurevich, Scott A. Waterman, Dick Crouch and Tracy Holloway King

A Web-interface for Eliciting Lexical Type of New Lexemes
Joshua Hou

12:10 – 13:30	Lunch

Session 2

13:30 – 14:00	Autosegmental representations in an HPSG of Hausa *Berthold Crysmann*
14:00 – 14:30	Construction of a German HPSG grammar from a detailed treebank *Bart Cramer and Yi Zhang*
14:30 – 15:00	Parenthetical Constructions - an Argument against Modularity *Eva Banik*
15:00 – 15:30	Using Artificially Generated Data to Evaluate Statistical Machine Translation *Manny Rayner, Paula Estrella, Pierrette Bouillon, Beth Ann Hockey and Yukie Nakao*
15:30 – 16:00	Break

Session 3

16:00 – 16:30	Using Large-scale Parser Output to Guide Grammar Development *Ascander Dost and Tracy Holloway King*
16:30 – 17:00	A generalized method for iterative error mining in parsing results *Daniël de Kok, Jianqiang Ma and Gertjan van Noord*
17:00 – 18:00	Discussion Session – Moderator: Joakim Nivre

WS3: Knowledge and Reasoning for Answering Questions (KRAQ)

13:50 – 14:00	Welcome Address
14:00 – 15:00	Invited Talk, Knowledge and Reasoning for Medical Question-Answering *Pierre Zweigenbaum*
15:00 – 15:30	The Development of a Question-Answering Services System for the Farmer through SMS: Query Analysis *Mukda Suktarachan, Patthrawan Rattanamanee and Asanee Kawtrakul*
15:30 – 16:00	Coffee Break
16:00 – 16:20	QAST: Question Answering System for ThaiWikipedia *Wittawat Jitkrittum, Choochart Haruechaiyasak and Thanaruk Theeramunkong*
16:20 – 16:40	Some Challenges in the Design of Comparative and Evaluative Question Answering Systems *Nathalie Lim, Patrick Saint-Dizier and Rachel Edita Roxas*
16:40 – 17:00	Addressing How-to Questions using a Spoken Dialogue System: a Viable Approach? *Silvia Quarteroni and Patrick Saint-Dizier*
17:00 – 17:30	Invited Talk, Question Answering: Reasoning, Mining, or Crowdsourcing? *Manfred Stede*

WS4: Language Generation and Summarisation (UCNLG+Sum)

Session 1: Sentence Compression and Revision

8:30 – 9:00 Unsupervised Induction of Sentence Compression Rules
Joao Cordeiro, Gael Dias and Pavel Brazdil

9:00 – 9:30 A Parse-and-Trim Approach with Information Significance for Chinese Sentence Compression
Wei Xu and Ralph Grishman

9:30 – 10:00 Syntax-Driven Sentence Revision for Broadcast News Summarization
Hideki Tanaka, Akinori Kinoshita, Takeshi Kobayakawa, Tadashi Kumano and Naoto Katoh

10:00 –10.30 Coffee Break

Session 2: Invited Talk / Content Selection

10:30 – 11:30 Query-focused Summarization Using Text-to-Text Generation: When Information Comes from Multilingual Sources
Kathleen R. McKeown

11:30 – 12:00 A Classification Algorithm for Predicting the Structure of Summaries
Horacio Saggion

12:00 – 12:30 Optimization-based Content Selection for Opinion Summarization
Jackie Chi Kit Cheung, Giuseppe Carenini and Raymond T. Ng

12:30 – 13:50 Lunch

Session 3: Evaluation

13:50 – 15:00 **GREC 2009 Shared Task Evaluation results session**

The GREC Main Subject Reference Generation Challenge 2009: Overview and Evaluation Results
Anja Belz, Eric Kow, Jette Viethen and Albert Gatt

The GREC Named Entity Generation Challenge 2009: Overview and Evaluation Results
Anja Belz, Eric Kow and Jette Viethen

ICSI-CRF: The Generation of References to the Main Subject and Named Entities Using Conditional Random Fields
Benoit Favre and Bernd Bohnet

UDel: Generating Referring Expressions Guided by Psycholinguistc Findings
Charles Greenbacker and Kathleen McCoy

JUNLG-MSR: A Machine Learning Approach of Main Subject Reference Selection with Rule Based Improvement
Samir Gupta and Sivaji Bandyopadhyay

UDel: Extending Reference Generation to Multiple Entities
Charles Greenbacker and Kathleen McCoy

WLV: A Confidence-based Machine Learning Method for the GREC-NEG'09 Task
Constatin Orasan and Iustin Dornescu

15:00 – 15:30 Evaluation of Automatic Summaries: Metrics under Varying Data Conditions
Karolina Owkzarzak and Hoa Trang Dang

15:30 – 16:00 Coffee Break

Session 4: Short Papers/Discussion

16:00 – 16:20 Visual Development Process for Automatic Generation of Digital Games Narrative Content
Maria Fernanda Caropreso, Diana Inkpen, Shahzad Khan and Fazel Keshtkar

16:20 – 16:40 Reducing Redundancy in Multi-document Summarization Using Lexical Semantic Similarity
Iris Hendrickx, Walter Daelemans, Erwin Marsi and Emiel Krahmer

16:40 – 17:00 Non-textual Event Summarization by Applying Machine Learning to Template-based Language Generation
Mohit Kumar, Dipanjan Das, Sachin Agarwal and Alexander I. Rudnicky

17:00 – 17:20 Creating an Annotated Corpus for Generating Walking Directions
Stephanie Schuldes, Michael Roth, Anette Frank and Michael Strube

17:20 – 18:00 Panel-led discussion on synergies between summarisation and NLG, including shared tasks

WS5: Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE)

8:30 – 8:45	Welcome and Introduction to the Workshop

Session 1 – MWE Identification and Disambiguation

8:45 – 9:10	Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains *Helena Caseli, Aline Villavicencio, André Machado and Maria José Finatto*
9:10 – 9:35	Re-examining Automatic Keyphrase Extraction Approaches in Scientific Articles *Su Nam Kim and Min-Yen Kan*
9:35 – 10:00	Verb Noun Construction MWE Token Classification *Mona T. Diab and Pravin Bhutada*

10:00 – 10:30	Break

Session 2 – Identification, Interpretation, and Disambiguation

10:30 – 10:55	Exploiting Translational Correspondences for Pattern-Independent MWE Identification *Sina Zarrieß and Jonas Kuhn*
10:55 – 11:20	A re-examination of lexical association measures *Hung Huu Hoang, Su Nam Kim and Min-Yen Kan*
11:20 – 11:45	Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus *R. Mahesh K. Sinha*

11:45 – 13:50	Lunch

Session 3 – Applications

13:50 – 14:15	Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions *Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu and Yun Huang*
14:15 – 14:40	Bottom-up Named Entity Recognition using Two-stage Machine Learning Method *Hirotaka Funayama, Tomohide Shibata and Sadao Kurohashi*
14:40 – 15:05	Abbreviation Generation for Japanese Multi-Word Expressions *Hiromi Wakaki, Hiroko Fujii, Masaru Suzuki, Mika Fukui and Kazuo Sumita*

15:05 – 15:30	Discussion of Sessions 1, 2, 3
15:30 – 16:00	Break
16:00 – 17:00	General Discussion
17:00 – 17:15	Closing Remarks

WS6: The Third Linguistic Annotation Workshop (LAW III)

8:50 – 9:00 Opening Remarks

9:00 – 9:30 A Cognitive-based Annotation System for Emotion Computing
Ying Chen, Sophia Y. M. Lee and Chu-Ren Huang

9:30 – 10:00 Complex Linguistic Annotation—No Easy Way Out! A Case from Bangla and Hindi POS Labeling Tasks
Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury and Kalika Bali

10:00 – 10:30 Break

10:30 – 11:00 Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation
Ines Rehbein, Josef Ruppenhofer and Caroline Sporleder

11:00 – 11:30 Bridging the Gaps: Interoperability for GrAF, GATE, and UIMA
Nancy Ide and Keith Suderman

11:30 – 12:00 By all these lovely tokens... Merging Conflicting Tokenizations
Christian Chiarcos, Julia Ritz and Manfred Stede

12:00 – 13:30 Lunch Break

13:30 – 13:50 Annotating Subordinators in the Turkish Discourse Bank
Deniz Zeyrek, Ümit Deniz Turan, Cem Bozsahin, Ruket Çakıcı, Ayışığı B. Sevdik-Çallı, Işın Demirşahin, Berfin Aktaş, İhsan Yalçınkaya and Hale Ögel

13:50 – 14:10 Annotation of Events and Temporal Expressions in French Texts
André Bittar

14:10 – 14:30 Designing a Language Game for Collecting Coreference Annotation
Barbora Hladká, Jiří Mírovský and Pavel Schlesinger

14:30 – 14:50 Explorations in Automatic Image Annotation using Textual Features
Chee Wee Leong and Rada Mihalcea

14:50 – 15:10	Human Evaluation of Article and Noun Number Usage: Influences of Context and Construction Variability *John Lee, Joel Tetreault and Martin Chodorow*
15:10 – 15:30	Stand-off TEI Annotation: the Case of the National Corpus of Polish *Piotr Banski and Adam Przepiórkowski*
15:30 – 16:00	Break
16:00 – 17:30	Poster Session
17:30 – 18:00	SIGANN Annual Meeting

WS7: 2nd Workshop on Building and Using Comparable Corpora: from parallel to non-parallel corpora (BUCC)

8:45 – 9:00	Welcome and Introduction
9:00 – 10:00	Invited Presentation: Repetition and Language Models and Comparable Corpora *Ken Church*
10:00 – 10:30	Coffee break

Session 1: Information Extraction and Summarization

10:30 – 10:55	Extracting Lay Paraphrases of Specialized Expressions from Monolingual Comparable Medical Corpora *Louise Deléger and Pierre Zweigenbaum*
10:55 – 11:20	An Extensible Crosslinguistic Readability Framework *Jesse Kirchner, Justin Nuger and Yi Zhang*
11:20 – 11:45	An Analysis of the Calque Phenomena Based on Comparable Corpora *Marie Garnier and Patrick Saint-Dizier*
11:45 – 12:10	Active Learning of Extractive Reference Summaries for Lecture Speech Summarization *Jian Zhang and Pascale Fung*
12:10 – 13:50	Lunch break

Session 2: Statistical Machine Translation

13:50 – 14:15	Train the Machine with What It Can Learn—Corpus Selection for SMT *Xiwu Han, Hanzhang Li and Tiejun Zhao*
14:15 – 14:40	Mining Name Translations from Comparable Corpora by Creating Bilingual Information Networks *Heng Ji*
14:40 – 15:05	Chinese-Uyghur Sentence Alignment: An Approach Based on Anchor Sentences *Samat Mamitimin and Min Hou*
15:05 – 15:30	Exploiting Comparable Corpora with TER and TERp *Sadaf Abdul Rauf and Holger Schwenk*
15:30 – 16:00	Coffee break

Session 3: Building Comparable Corpora

16:00 – 16:25	Compilation of Specialized Comparable Corpora in French and Japanese *Lorraine Goeuriot, Emmanuel Morin and Béatrice Daille*
16:25 – 16:50	Toward Categorization of Sign Language Corpora *Jérémie Segouat and Annelies Braffort*
16:50 – 17:50	Panel Session – Multilingual Information Processing: from Parallel to Comparable Corpora

WS10: The 7th Workshop on Asian Language Resources (ALR7)

8:30 – 9:00	Registration
9:00 – 9:10	Opening
9:10 – 9:35	Enhancing the Japanese WordNet *Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki*
9:35 – 10:00	An Empirical Study of Vietnamese Noun Phrase Chunking with Discriminative Sequence Models *Le Minh Nguyen, Huong Thao Nguyen, Phuong Thai Nguyen, Tu Bao Ho and Akira Shimazu*
10:00 – 10:30	Break
10:30 – 10:55	Corpus-based Sinhala Lexicon *Ruvan Weerasinghe, Dulip Herath and Viraj Welgama*
10:55 – 11:20	Analysis and Development of Urdu POS Tagged Corpus *Ahmed Muaz, Aasim Ali and Sarmad Hussain*
11:20 – 11:45	Annotating Dialogue Acts to Construct Dialogue Systems for Consulting *Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka and Satoshi Nakamura*
11:45 – 12:10	Assas-band, an Affix-Exception-List Based Urdu Stemmer *Qurat-ul-Ain Akram, Asma Naseer and Sarmad Hussain*
12:10 – 13:50	Lunch break
13:50 – 14:15	Automated Mining Of Names Using Parallel Hindi-English Corpus *R. Mahesh K. Sinha*
14:15 – 14:40	Basic Language Resources for Diverse Asian Languages: A Streamlined Approach for Resource Creation *Heather Simpson, Kazuaki Maeda and Christopher Cieri*
14:40 – 15:05	Finite-State Description of Vietnamese Reduplication *Le Hong Phuong, Nguyen Thi Minh Huyen and Roussanaly Azim*
15:05 – 15:30	Construction of Chinese Segmented and POS-tagged Conversational Corpora and Their Evaluations on Spontaneous Speech Recognitions *Xinhui Hu, Ryosuke Isotani and Satoshi Nakamura*

15:30 – 16:00	Break
16:00 – 16:15	Bengali Verb Subcategorization Frame Acquisition - A Baseline Model *Somnath Banerjee, Dipankar Das and Sivaji Bandyopadhyay*
16:15 – 16:30	Phonological and Logographic Influences on Errors in Written Chinese Words *Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang and Shih-Hung Wu*
16:30 – 16:45	Resource Report: Building Parallel Text Corpora for Multi-Domain Translation System *Budiono, Hammam Riza and Chairil Hakim*
16:45 – 17:00	A Syntactic Resource for Thai: CG Treebank *Taneth Ruangrajitpakorn, Kanokorn Trakultaweekoon and Thepchai Supnithi*
17:00 – 17:15	Part of Speech Tagging for Mongolian Corpus *Purev Jaimai and Odbayar Chimeddorj*

10

Friday, 7 August

Session	EMNLP Oral A	EMNLP Oral B	EMNLP Oral C	EMNLP Oral D	Workshops
Venue	Theater	MR208	MR209	MR203	Various MR, Level 3
8:30-10:00	4A: Multiword Expressions	4B: Machine Learning and Statistical Models II	4C: Information Retrieval and Question Answering	4D: Syntax and Parsing I	Morning Session 1
10:00-10:30	Coffee/Tea Break				
10:30-12:10	5A: Subjectivity and Sentiment II	5B: Lexical Sematics II	5C: Phonology and Morphology	5D: Machine Translation IV	Morning Session 2
12:10-13:50	Lunch				
13:50-15:30	6A: Speech and Language Modeling	6B: Semantic Similarity	6C: Syntax and Parsing II	6D: Multilinguality	Afternoon Session 2
15:30-16:00	Coffee/Tea Break				
16:00-18:00	7A: Natural Language Applications	7B: Lexical Semantics III	7C: Coreference Resolution	7D: Machine Translation V	Afternoon Session 2

EMNLP – Day 2

Session 4A (Theatre): Multi-word Expressions

8:45 – 9:10	Multi-Word Expression Identification Using Sentence Surface Features *Ram Boukobza and Ari Rappoport*
9:10 – 9:35	Acquiring Translation Equivalences of Multiword Expressions by Normalized Correlation Frequencies *Ming-Hong Bai, Jia-Ming You, Keh-Jiann Chen and Jason S. Chang*
9:35 – 10:00	Collocation Extraction Using Monolingual Word Alignment Method *Zhanyi Liu, Haifeng Wang, Hua Wu and Sheng Li*

Session 4B (MR208): Machine Learning and Statistical Models II

8:45 – 9:10	Multi-Class Confidence Weighted Algorithms *Koby Crammer, Mark Dredze and Alex Kulesza*
9:10 – 9:35	Model Adaptation via Model Interpolation and Boosting for Web Search Ranking *Jianfeng Gao, Qiang Wu, Chris Burges, Krysta Svore, Yi Su, Nazan Khan, Shalin Shah and Hongyan Zhou*
9:35 – 10:00	A Structural Support Vector Method for Extracting Contexts and Answers of Questions from Online Forums *Wen-Yun Yang, Yunbo Cao and Chin-Yew Lin*

Session 4C (MR209): Information Retrieval and Questions Answering

8:45 – 9:10	Mining Search Engine Clickthrough Log for Matching N-gram Features *Huihsin Tseng, Longbin Chen, Fan Li, Ziming Zhuang, Lei Duan and Belle Tseng*
9:10 – 9:35	The role of named entities in Web People Search *Javier Artiles, Enrique Amigó and Julio Gonzalo*
9:35 – 10:00	Investigation of Question Classifier in Question Answering *Zhiheng Huang, Marcus Thint and Asli Celikyilmaz*

Session 4D (MR203): Syntax and Parsing I

8:45 – 9:10	An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing *Jun Suzuki, Hideki Isozaki, Xavier Carreras and Michael Collins*
9:10 – 9:35	Statistical Bistratal Dependency Parsing *Richard Johansson*
9:35 – 10:00	Improving Dependency Parsing with Subtrees from Auto-Parsed Data *Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto and Kentaro Torisawa*

10:00 – 10:30	Coffee Break

Session 5A (Theatre): Subjectivity and Sentiment II

10:30 – 10:55	Topic-wise, Sentiment-wise, or Otherwise? Identifying the Hidden Dimension for Unsupervised Text Classification *Sajib Dasgupta and Vincent Ng*
10:55 – 11:20	Adapting a Polarity Lexicon using Integer Linear Programming for Domain-Specific Sentiment Classification *Yejin Choi and Claire Cardie*
11:20 – 11:45	Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus *Saif Mohammad, Cody Dunne and Bonnie Dorr*
11:45 – 12:10	Matching Reviews to Objects using a Language Model *Nilesh Dalvi, Ravi Kumar, Bo Pang and Andrew Tomkins*

Session 5B (MR208): Lexical Semantics II

10:30 – 10:55	EEG responds to conceptual stimuli and corpus semantics *Brian Murphy, Marco Baroni and Massimo Poesio*
10:55 – 11:20	A Comparison of Windowless and Window-Based Computational Association Measures as Predictors of Syntagmatic Human Associations *Justin Washtell and Katja Markert*

11:20 – 11:45 Improving Verb Clustering with Automatically Acquired Selectional Preferences
Lin Sun and Anna Korhonen

11:45 – 12:10 Improving Web Search Relevance with Semantic Features
Yumao Lu, Fuchun Peng, Gilad Mishne, Xing Wei and Benoit Dumoulin

Session 5C (MR209): Phonology and Morphology

10:30 – 10:55 Can Chinese Phonemes Improve Machine Transliteration?: A Comparative Study of English-to-Chinese Transliteration Models
Jong-Hoon Oh, Kiyotaka Uchimoto and Kentaro Torisawa

10:55 – 11:20 Unsupervised morphological segmentation and clustering with document boundaries
Taesun Moon, Katrin Erk and Jason Baldridge

11:20 – 11:45 The infinite HMM for unsupervised PoS tagging
Jurgen Van Gael, Andreas Vlachos and Zoubin Ghahramani

11:45 – 12:10 A Simple Unsupervised Learner for POS Disambiguation Rules Given Only a Minimal Lexicon
Qiuye Zhao and Mitch Marcus

Session 5D (MR203): Machine Translation IV

10:30 – 10:55 Tree Kernel-based SVM with Structured Syntactic Knowledge for BTG-based Phrase Reordering
Min Zhang and Haizhou Li

10:55 – 11:20 Discriminative Corpus Weight Estimation for Machine Translation
Spyros Matsoukas, Antti-Veikko I. Rosti and Bing Zhang

11:20 – 11:45 Unsupervised Tokenization for Machine Translation
Tagyoung Chung and Daniel Gildea

11:45 – 12:10 Synchronous Tree Adjoining Machine Translation
Steve DeNeefe and Kevin Knight

12:10 – 13:50 Lunch

Session 6A (Theatre): Speech and Language Modeling

13:50 – 14:15	Word Buffering Models for Improved Speech Repair Parsing *Tim Miller*
14:15 – 14:40	Less is More: Significance-Based N-gram Selection for Smaller, Better Language Models *Robert C. Moore and Chris Quirk*
14:40 – 15:05	Stream-based Randomised Language Models for SMT *Abby Levenberg and Miles Osborne*
15:05 – 15:30	Integrating sentence- and word-level error identification for disfluency correction *Erin Fitzgerald, Frederick Jelinek and Keith Hall*

Session 6B (MR208): Semantic Similarity

13:50 – 14:15	Estimating Semantic Distance Using Soft Semantic Constraints in Knowledge-Source . Corpus Hybrid Models *Yuval Marton, Saif Mohammad and Philip Resnik*
14:15 – 14:40	Recognizing Textual Relatedness with Predicate-Argument Structures *Rui Wang and Yi Zhang*
14:40 – 15:05	Learning Term-weighting Functions for Similarity Measures *Wen-tau Yih*
15:05 – 15:30	A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web *Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka*

Session 6C (MR209): Syntax and Parsing II

13:50 – 14:15	Unbounded Dependency Recovery for Parser Evaluation *Laura Rimell, Stephen Clark and Mark Steedman*
14:15 – 14:40	Parser Adaptation and Projection with Quasi-Synchronous Grammar Features *David A. Smith and Jason Eisner*
14:40 – 15:05	Self-Training PCFG Grammars with Latent Annotations Across Languages *Zhongqiang Huang and Mary Harper*
15:05 – 15:30	An Alternative to Head-Driven Approaches for Parsing a (Relatively) Free Word-Order Language *Reut Tsarfaty, Khalil Sima'an and Remko Scha*

Session 6D (MR203): Multilinguality

13:50 – 14:15	Enhancement of Lexical Concepts Using Cross-lingual Web Mining *Dmitry Davidov and Ari Rappoport*
14:15 – 14:40	Bilingual dictionary generation for low-resourced language pairs *István Varga and Shoichi Yokoyama*
14:40 – 15:05	Multilingual Spectral Clustering Using Document Similarity Propagation *Dani Yogatama and Kumiko Tanaka-Ishii*
15:05 – 15:30	Polylingual Topic Models *David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith and Andrew McCallum*
15:30 – 16:00	Coffee Break

Session 7A (Theatre): Natural Language Applications

16:00 – 16:25	Using the Web for Language Independent Spellchecking and Autocorrection *Casey Whitelaw, Ben Hutchinson, Grace Y. Chung and Ged Ellis*
16:25 – 16:50	Statistical Estimation of Word Acquisition with Application to Readability Prediction *Paul Kidwell, Guy Lebanon and Kevyn Collins-Thompson*
16:50 – 17:15	Combining Collocations, Lexical and Encyclopedic Knowledge for Metonymy Resolution *Vivi Nastase and Michael Strube*
17:15 – 17:40	Segmenting Email Message Text into Zones *Andrew Lampert, Robert Dale and Cécile Paris*

Session 7B (MR208): Lexical Semantics III

16:00 – 16:25	Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures *Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond and Asuka Sumida*
16:25 – 16:50	Web-Scale Distributional Similarity and Entity Set Expansion *Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu and Vishnu Vyas*

16:50 – 17:15 Toward Completeness in Concept Extraction and Classification
Eduard Hovy, Zornitsa Kozareva and Ellen Riloff

17:15 – 17:40 Reading to Learn: Constructing Features from Semantic Abstracts
Jacob Eisenstein, James Clarke, Dan Goldwasser and Dan Roth

Session 7C (MR209): Coreference Resolution

16:00 – 16:25 Supervised Models for Coreference Resolution
Altaf Rahman and Vincent Ng

16:25 – 16:50 Global Learning of Noun Phrase Anaphoricity in Coreference Resolution via Label Propagation
GuoDong Zhou and Fang Kong

16:50 – 17:15 Employing the Centering Theory in Pronoun Resolution from the Semantic Perspective
Fang Kong, GuoDong Zhou and Qiaoming Zhu

17:15 – 17:40 Person Cross Document Coreference with Name Perplexity Estimates
Octavian Popescu

Session 7D (MR203): Machine Translation V

16:00 – 16:25 Learning Linear Ordering Problems for Better Translation
Roy Tromble and Jason Eisner

16:25 – 16:50 Weighted Alignment Matrices for Statistical Machine Translation
Yang Liu, Tian Xia, Xinyan Xiao and Qun Liu

16:50 – 17:15 Sinuhe . Statistical Machine Translation using a Globally Trained Conditional Exponential Family Translation Model
Matti Kääriäinen

17:15 – 17:40 Fast Translation Rule Matching for Syntax-based Statistical Machine Translation
Hui Zhang, Min Zhang, Haizhou Li and Chew Lim Tan

Workshop Listing

6-7 August – 2 day workshops

WS6: The Third Linguistic Annotation Workshop (LAW III)
Venue: MR303
Chairs: Nancy Ide, Adam Meyers, Manfred Stede, Chu-Ren Huang

WS10: The 7th Workshop on Asian Language Resources (ALR7)
Venue: MR302
Chairs: Hammam Riza, Virach Sornlertlamvanich

7 August – 1 day workshops

WS8: TextGraphs-4: Graph-based Methods for Natural Language Processing
Venue: MR301
Chairs: Monojit Choudhury, Samer Hassan, Animesh Mukherjee, Smaranda Muresan

WS9: The People's Web meets NLP: Collaboratively Constructed Semantic Resources (People's Web)
Venue: MR304
Chairs: Iryna Gurevych, Torsten Zesch

WS11: Named Entities Workshop - Shared Task on Transliteration (NEWS on Transliteration)
Venue: MR305
Chairs: Haizhou Li, A Kumaran

WS12: Workshop on text and citation analysis for scholarly digital libraries (NLPIR4DL)
Venue: MR306
Chairs: Simone Teufel, Min-Yen Kan

WS6: The Third Linguistic Annotation Workshop (LAW III)

9:00 – 9:30	Committed Belief Annotation and Tagging *Mona T. Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaram and Weiwei Guo*
9:30 – 10:00	Annotation of Sentence Structure: Capturing the Relationship among Clauses in Czech Sentences *Markéta Lopatková, Natalia Klyueva and Petr Homola*
10:00 – 10:30	Break
10:30 – 11:00	Schema and Variation: Digitizing Printed Dictionaries *Christian Schneiker, Dietmar Seipel and Werner Wegstein*
11:00 – 11:30	Syntactic Annotation of Spoken Utterances: A Case Study on the Czech Academic Corpus *Barbora Hladká and Zdenka Uresova*
11:30 – 12:00	High-Performance High-Volume Layered Corpora Annotation *Tiago Luís and David Martins de Matos*
12:00 – 13:30	Lunch Break
13:30 – 13:50	The Coding Scheme for Annotating Extended Nominal Coreference and Bridging Anaphora in the Prague Dependency Treebank *Anna Nedoluzhko, Jiří Mírovský and Petr Pajas*
13:50 – 14:10	Timed Annotations — Enhancing MUC7 Metadata by the Time It Takes to Annotate Named Entities *Katrin Tomanek and Udo Hahn*
14:10 – 14:30	Transducing Logical Relations from Automatic and Manual GLARF *Adam Meyers, Michiko Kosaka, Heng Ji, Nianwen Xue, Mary Harper, Ang Sun, Wei Xu and Shasha Liao*
14:30 – 14:50	Using Parallel Propbanks to enhance Word-alignments *Jinho Choi, Martha Palmer and Nianwen Xue*
14:50 – 15:10	WordNet and FrameNet as Complementary Resources for Annotation *Collin F. Baker and Christiane Fellbaum*
15:10 – 15:30	SIGANN Working Group Reports

15:30 – 16:00	Break
16:00 – 17:50	Panel: The Standards Debate: Pro and Con
17:50 – 18:00	Closing

WS8: TextGraphs-4: Graph-based Methods for Natural Language Processing

Session I: Opening

8:30 – 8:45	Inauguration by Chairs
8:45 – 9:48	Invited Talk *Vittorio Loreto*
9:48 – 10:00	Social (distributed) language modeling, clustering and dialectometry *David Ellis*
10:00 – 10:30	Coffee Break

Session II: Special Theme

10:30 – 10:55	Network analysis reveals structure indicative of syntax in the corpus of undeciphered Indus civilization inscriptions *Sitabhra Sinha, Raj Kumar Pan, Nisha Yadav, Mayank Vahia and Iravatham Mahadevan*
10:55 – 11:20	Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology *Martijn Wieling and John Nerbonne*
11:20 – 12:10	Panel Discussion

Session III: Semantics

13:50 – 14:15	Random Walks for Text Semantic Similarity *Daniel Ramage, Anna N. Rafferty and Christopher D. Manning*
14:15 – 14:40	Classifying Japanese Polysemous Verbs based on Fuzzy C-means Clustering *Yoshimi Suzuki and Fumiyo Fukumoto*
14:40 – 15:05	WikiWalk: Random walks on Wikipedia for Semantic Relatedness *Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre and Aitor Soroa*
15:05 – 15:18	Measuring semantic relatedness with vector space models and random walks *Amaç Herdağdelen, Katrin Erk and Marco Baroni*
15:18 – 15:30	Graph-based Event Coreference Resolution *Zheng Chen and Heng Ji*
15:30 – 16:00	Coffee Break

Session IV: Classification and Clustering

16:00 – 16:25	Ranking and Semi-supervised Classification on Large Scale Graphs Using Map-Reduce *Delip Rao and David Yarowsky*
16:25 – 16:50	Opinion Graphs for Polarity and Discourse Classification *Swapna Somasundaran, Galileo Namata, Lise Getoor and Janyce Wiebe*
16:50 – 17:15	A Cohesion Graph Based Approach for Unsupervised Recognition of Literal and Non-literal Use of Multiword Expressions *Linlin Li and Caroline Sporleder*
17:15 – 17:40	Quantitative analysis of treebanks using frequent subtree mining methods *Scott Martens*
17:40 – 18:00	Closing

WS9: The People's Web meets NLP: Collaboratively Constructed Semantic Resources (People's Web)

8:45 – 9:00	Opening Remarks
9:00 – 9:30	A Novel Approach to Automatic Gazetteer Generation using Wikipedia *Ziqi Zhang and Jose Iria*
9:30 – 10:00	Named Entity Recognition in Wikipedia *Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy and James R. Curran*
10:00 – 10:30	Coffee Break
10:30 – 11:00	Wiktionary for Natural Language Processing: Methodology and Limitations *Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, Shu-Kai Hsieh, Ivy Kuo, Pierre Magistry and Chu-Ren Huang*
11:00 – 11:20	Using the Wiktionary Graph Structure for Synonym Detection *Timothy Weale, Chris Brew and Eric Fosler-Lussier*
11:20 – 11:40	Automatic Content-Based Categorization of Wikipedia Articles *Zeno Gantner and Lars Schmidt-Thieme*
11:40 – 12:00	Evaluating a Statistical CCG Parser on Wikipedia *Matthew Honnibal, Joel Nothman and James R. Curran*
12:10 – 13:50	Lunch Break
13:50 – 14:50	Invited Talk *Rada Mihalcea*
15:00 – 15:30	Construction of Disambiguated Folksonomy Ontologies Using Wikipedia *Noriko Tomuro and Andriy Shepitsen*
15:30 – 16:00	Coffee Break
16:00 – 16:20	Acquiring High Quality Non-Expert Knowledge from On-Demand Workforce *Donghui Feng, Sveva Besana and Remi Zajac*

16:20 – 16:40	Constructing an Anaphorically Annotated Corpus with Non-Experts: Assessing the Quality of Collaborative Annotations *Jon Chamberlain, Udo Kruschwitz and Massimo Poesio*
16:40 – 17:00	Discussion

WS10: The 7th Workshop on Asian Language Resources (ALR7)

8:45 – 9:10 Interaction Grammar for the Persian Language: Noun and Adjectival Phrases
Masood Ghayoomi and Bruno Guillaume

9:10 – 9:35 KTimeML: Specification of Temporal and Event Expressions in Korean Text
Seohyun Im, Hyunjo You, Hayun Jang, Seungho Nam and Hyopil Shin

9:35 – 10:00 CWN-LMF: Chinese WordNet in the Lexical Markup Framework
Lung-Hao Lee, Shu-Kai Hsieh and Chu-Ren Huang

10:00 – 10:30 Break

10:30 – 10:55 Philippine Language Resources: Trends and Directions
Rachel Edita Roxas, Charibeth Cheng and Nathalie Rose Lim

10:55 – 11:20 Thai WordNet Constructio
Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich and Hitoshi Isahara

11:20 – 11:45 Query Expansion using LMF-Compliant Lexical Resources
Takenobu Tokunaga, Dain Kaplan, Nicoletta Calzolari, Mon Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Yingju Xia, Chu-Ren Huang, Shu-Kai Hsieh and Kiyoaki Shirai

11:45 – 12:10 Thai National Corpus: A Progress Report
Wirote Aroonmanakun, Kachen Tansiri and Pairit Nittayanuparp

12:10 – 13:50 Lunch break

13:50 – 14:15 The FLaReNet Thematic Network: A Global Forum for Cooperation
Nicoletta Calzolari and Claudia Soria

14:15 – 14:40 Towards Building Advanced Natural Language Applications - An Overview of the Existing Primary Resources and Applications in Nepali
Bal Krishna Bal

14:40 – 15:05	Using Search Engine to Construct a Scalable Corpus for Vietnamese Lexical Development for Word Segmentation *Doan Nguyen*
15:05 – 15:30	Word Segmentation Standard in Chinese, Japanese and Korean *Key-Sun Choi, Hitoshi Isahara, Kyoko Kanzaki, Hansaem Kim, Seok Mun Pak and Maosong Sun*
15:30 – 16:00	Break
16:00 – 17:50	Panel discussion "ALR and FLaReNet"
17:50 – 18:00	Closing

WS11: Named Entities Workshop - Shared Task on Transliteration (NEWS on Transliteration)

8:30 – 9:10 Opening Remarks – Overview of the Shared Tasks – Report of NEWS 2009 Machine Transliteration Shared Task
Haizhou Li, A Kumaran, Vladimir Pervouchine and Min Zhang

9:10 – 10:00 Keynote Speech – Automata for Transliteration and Machine Translation
Kevin Knight

10:00 – 10:30 Coffee Break

Session 1: Shared Task Paper Presentation

10:30 – 10:45 DirecTL: a Language Independent Approach to Transliteration
Sittichai Jiampojamarn, Aditya Bhargava, Qing Dou, Kenneth Dwyer and Grzegorz Kondrak

10:45 – 11:00 Named Entity Transcription with Pair n-Gram Models
Martin Jansche and Richard Sproat

11:00 – 11:15 Machine Transliteration using Target-Language Grapheme and Phoneme: Multi-engine Transliteration Approach
Jong-Hoon Oh, Kiyotaka Uchimoto and Kentaro Torisawa

11:15 – 11:30 A Language-Independent Transliteration Schema Using Character Aligned Models at NEWS 2009
Praneeth Shishtla, Surya Ganesh V, Sethuramalingam Subramaniam and Vasudeva Varma

11:30 – 11:45 Experiences with English-Hindi, English-Tamil and English-Kannada Transliteration Tasks at NEWS 2009
Manoj Kumar Chinnakotla and Om P. Damani

12:00 – 13:50 Lunch Break

13:50 – 15:30 **Session 2: Poster Presentations**

Testing and Performance Evaluation of Machine Transliteration System for Tamil Language
Kommaluri Vijayanand

13:50 – 15:30 **Session 2: Poster Presentations (con't)**

Transliteration by Bidirectional Statistical Machine Translation
Andrew Finch and Eiichiro Sumita

Transliteration of Name Entity via Improved Statistical Translation on Character Sequences
Yan Song, Chunyu Kit and Xiao Chen

Learning Multi Character Alignment Rules and Classification of Training Data for Transliteration
Dipankar Bose and Sudeshna Sarkar

Fast Decoding and Easy Implementation: Transliteration as Sequential Labeling
Eiji Aramaki and Takeshi Abekawa

NEWS 2009 Machine Transliteration Shared Task System Description: Transliteration with Letter-to-Phoneme Technology
Colin Cherry and Hisami Suzuki

Combining a Two-step Conditional Random Field Model and a Joint Source Channel Model for Machine Transliteration
Dong Yang, Paul Dixon, Yi-Cheng Pan, Tasuku Oonishi, Masanobu Nakamura and Sadaoki Furui

Phonological Context Approximation and Homophone Treatment for NEWS 2009 English-Chinese Transliteration Shared Task
Oi Yee Kwong

English to Hindi Machine Transliteration System at NEWS 2009
Amitava Das, Asif Ekbal, Tapabrata Mondal and Sivaji Bandyopadhyay

Improving Transliteration Accuracy Using Word-Origin Detection and Lexicon Lookup
Mitesh M. Khapra and Pushpak Bhattacharyya

A Noisy Channel Model for Grapheme-based Machine Transliteration
Jia Yuxiang, Zhu Danqing and Yu Shiwen

Substring-based Transliteration with Conditional Random Fields
Sravana Reddy and Sonjia Waxmonsky

A Syllable-based Name Transliteration System
Xue Jiang, Le Sun and Dakun Zhang

Transliteration System Using Pair HMM with Weighted FSTs
Peter Nabende

English-Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009
Rejwanul Haque, Sandipan Dandapat, Ankit Kumar Srivastava, Sudip Kumar Naskar and Andy Way

A Hybrid Approach to English-Korean Name Transliteration
Gumwon Hong, Min-Jeong Kim, Do-Gil Lee and Hae-Chang Rim

13:50 – 15:30 **Session 2: Poster Presentations (con't)**

Language Independent Transliteration System Using Phrase-based SMT Approach on Substrings
Sara Noeman

Combining MDL Transliteration Training with Discriminative Modeling
Dmitry Zelenko

ϵ-extension Hidden Markov Models and Weighted Transducers for Machine Transliteration
Balakrishnan Varadarajan and Delip Rao

Modeling Machine Transliteration as a Phrase Based Statistical Machine Translation Problem
Taraka Rama and Karthik Gali

Maximum n-Gram HMM-based Name Transliteration: Experiment in NEWS 2009 on English-Chinese Corpus
Yilu Zhou

Name Transliteration with Bidirectional Perceptron Edit Models
Dayne Freitag and Zhiqiang Wang

Bridging Languages by SuperSense Entity Tagging
Davide Picca, Alfio Massimiliano Gliozzo and Simone Campora

Chinese-English Organization Name Translation Based on Correlative Expansion
Feiliang Ren, Muhua Zhu, Huizhen Wang and Jingbo Zhu

Name Matching between Roman and Chinese Scripts: Machine Complements Human
Ken Samuel, Alan Rubenstein, Sherri Condon and Alex Yeh

Analysis and Robust Extraction of Changing Named Entities
Masatoshi Tsuchiya, Shoko Endo and Seiichi Nakagawa

15:30 – 16:00 Tag Confidence Measure for Semi-Automatically Updating Named Entity Recognition
Kuniko Saito and Kenji Imamura

16:20 – 16:40 A Hybrid Model for Urdu Hindi Transliteration
Abbas Malik, Laurent Besacier, Christian Boitet and Pushpak Bhattacharyya

16:40 – 17:00 Graphemic Approximation of Phonological Context for English-Chinese Transliteration
Oi Yee Kwong

17:00 – 17:20 Czech Named Entity Corpus and SVM-based Recognizer
Jana Kravalova and Zdenek Zabokrtsky

17:20 – 17:40 Voted NER System using Appropriate Unlabeled Data
Asif Ekbal and Sivaji Bandyopadhyay

WS12: Workshop on text and citation analysis for scholarly digital libraries (NLPIR4DL)

9:00 – 9:10	Opening Remarks
9:10 – 10:00	Invited Talk by Rick Lee of the World Scientific Publishing Company
10:00 – 10:30	Coffee Break

Session 1: Metadata and Content

10:30 – 10:55	Researcher affiliation extraction from homepages *István Nagy, Richárd Farkas and Márk Jelasity*
10:55 – 11:20	Anchor Text Extraction for Academic Search *Shuming Shi, Fei Xing, Mingjie Zhu, Zaiqing Nie and Ji-Rong Wen*
11:20 – 11:45	Accurate Argumentative Zoning with Maximum Entropy models *Stephen Merity, Tara Murphy and James R. Curran*
11:45 – 12:10	Classification of Research Papers into a Patent Classification System Using Two Translation Models *Hidetsugu Nanba and Toshiyuki Takezawa*
12:10 – 13:50	Lunch Break

Session 2: System Aspects

13:50 – 14:15	Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences *Ágnes Sándor and Angela Vorndran*
14:15 – 14:40	Designing a Citation-Sensitive Research Tool: An Initial Study of Browsing-Specific Information Needs *Stephen Wan, Cécile Paris, Michael Muthukrishna and Robert Dale*
14:40 – 15:05	The ACL Anthology Network *Dragomir R. Radev, Pradeep Muthukrishnan and Vahed Qazvinian*
15:05 – 15:30	NLP Support for Faceted Navigation in Scholarly Collection *Marti A. Hearst and Emilia Stoica*
15:30 – 16:00	Coffee Break

Session 3: Citation Support

16:00 – 16:25	FireCite: Lightweight real-time reference string extraction from webpages *Ching Hoi Andy Hong, Jesse Prabawa Gozali and Min-Yen Kan*
16:25 – 16:50	Citations in the Digital Library of Classics: Extracting Canonical References by Using Conditional Random Fields *Matteo Romanello, Federico Boschetti and Gregory Crane*
16:50 – 17:15	Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-chain based Approach *Dain Kaplan, Ryu Iida and Takenobu Tokunaga*
17:15 – 18:00	Informal Demonstration Sessions - Wrap up

11

ACL-IJCNLP Abstracts

Full Paper Presentations – Monday, 3 August

Investigations on Word Senses and Word Usages

Katrin Erk, Diana McCarthy and Nicholas Gaylord

The vast majority of work on word senses has relied on predefined sense inventories and an annotation schema where each word instance is tagged with the best fitting sense. This paper examines the case for a graded notion of word meaning in two experiments, one which uses WordNet senses in a graded fashion, contrasted with the “winner takes all” annotation, and one which asks annotators to judge the similarity of two usages. We find that the graded responses correlate with annotations from previous datasets, but sense assignments are used in a way that weakens the case for clear cut sense boundaries. The responses from both experiments correlate with the overlap of paraphrases from the English lexical substitution task which bodes well for the use of substitutes as a proxy for word sense. This paper also provides two novel datasets which can be used for evaluating computational systems.

A Comparative Study on Generalization of Semantic Roles in FrameNet

Yuichiroh Matsubayashi, Naoaki Okazaki and Jun’ichi Tsujii

A number of studies have presented machine-learning approaches to semantic role labeling with availability of corpora such as FrameNet and PropBank. These corpora define the semantic roles of predicates for each frame independently. Thus, it is crucial for the machine-learning approach to generalize semantic roles across different frames, and to increase the size of training instances. This paper explores several criteria for generalizing semantic roles in FrameNet: role hierarchy, human-understandable descriptors of roles, semantic types of filler phrases, and mappings from FrameNet roles to thematic roles of VerbNet. We also propose feature functions that naturally combine and weight these criteria, based on the training data. The experimental result of the role classification shows 19.16% and 7.42% improvements in error reduction rate and macro-averaged F1 score, respectively. We also provide indepth analyses of the proposed criteria.

Unsupervised Argument Identification for Semantic Role Labeling

Omri Abend, Roi Reichart and Ari Rappoport

The task of Semantic Role Labeling (SRL) is often divided into two sub-tasks: verb argument identification, and argument classification. Current SRL algorithms show lower results on the identification sub-task. Moreover, most SRL algorithms are supervised, relying on large amounts of manually created data. In this paper we present an unsupervised algorithm for identifying verb arguments, where the only type of annotation required is POS tagging. The algorithm makes use of a fully unsupervised syntactic parser, using its output in order to detect clauses and gather candidate argument collocation statistics. We evaluate our algorithm on PropBank10, achieving a precision of 56%, as opposed to 47% of a strong baseline. We also obtain an 8% increase in precision for a Spanish corpus. This is the first paper that tackles unsupervised verb argument identification without using manually encoded rules or extensive lexical or syntactic resources.

Brutus: A Semantic Role Labeling System Incorporating CCG, CFG, and Dependency Features

Stephen Boxwell, Dennis Mehay and Chris Brew

We describe a semantic role labeling system that makes primary use of CCG-based features. Most previously developed systems are CFG-based and make extensive use of a treepath feature, which often suffers from data sparsity due to its use of explicit tree configurations. CCG affords ways to augment treepath-based features to overcome these data sparsity issues. By adding features over CCG word-word dependencies and lexicalized verbal subcategorization frames ("supertags"), we can obtain an F-score that is substantially better than a previous CCG-based SRL system and competitive with the current state of the art. A manual error analysis reveals that parser errors account for many of the errors of our system. This analysis also suggests that simultaneous incremental parsing and semantic role labeling may lead to performance gains in both tasks.

Exploiting Heterogeneous Treebanks for Parsing

Zheng-Yu Niu, Haifeng Wang and Hua Wu

We address the issue of using heterogeneous treebanks for parsing by breaking it down into two sub-problems, converting grammar formalisms of the treebanks to the same one, and parsing on these homogeneous treebanks. First we propose to employ an iteratively trained target grammar parser to perform grammar formalism conversion, eliminating predefined heuristic rules as required in previous methods. Then we provide two strategies to refine conversion results, and adopt a corpus weighting technique for parsing on homogeneous treebanks. Results on the Penn Treebank show that our conversion method achieves 42% error reduction over the previous best result. Evaluation on the Penn Chinese Treebank indicates that a converted dependency treebank helps constituency parsing and the use of unlabeled data by self-training further increases parsing f-score to 85.2%, resulting in 6% error reduction over the previous best result.

Cross Language Dependency Parsing using a Bilingual Lexicon

Hai Zhao, Yan Song, Chunyu Kit and Guodong Guodong

This paper proposes an approach to enhance dependency parsing in a language by using a translated treebank from another language. A simple statistical machine translation method, word-by-word decoding, where not a parallel corpus but a bilingual lexicon is necessary, is adopted for the treebank translation. Using an ensemble method, the key information extracted from word pairs with dependency relations in the translated text is effectively integrated into the parser for the target language. The proposed method is evaluated in English and Chinese treebanks. It is shown that a translated English treebank helps a Chinese parser obtain a state-of-the-art result.

Topological Field Parsing of German

Jackie Chi Kit Cheung and Gerald Penn

Freer-word-order languages such as German exhibit linguistic phenomena that present unique challenges to traditional CFG parsing. Such phenomena produce discontinuous constituents, which are not naturally modelled by projective phrase structure trees. In this paper, we examine topological field parsing, a shallow form of parsing which identifies the major sections of a sentence in relation to the clausal main verb and the subordinating heads. We report the results of topological field parsing of German using the unlexicalized, latent variable-based Berkeley parser (Petrov et al., 2006) Without any language- or model-dependent adaptation, we achieve state-of-the-art results on the TüBa-D/Z corpus, and a modified NEGRA corpus that has been automatically annotated with topological fields (Becker and Frank, 2002). We also perform a qualitative error analysis of the parser output, and discuss strategies to further improve the parsing results.

Unsupervised Multilingual Grammar Induction

Benjamin Snyder, Tahira Naseem and Regina Barzilay

We investigate the task of unsupervised grammar induction from bilingual parallel corpora. Our goal is to use bilingual cues to learn improved parsing models for each language and to evaluate these models on held-out monolingual test data. We formulate a generative Bayesian model which seeks to explain the observed parallel data through a combination of bilingual and monolingual parameters. To this end, we adapt a formalism known as unordered tree alignment to our probabilistic setting. Using this formalism, our model loosely binds parallel trees while allowing language-specific syntactic structure. We perform inference under this model using Markov Chain Monte Carlo and dynamic programming. Applying this model to three parallel corpora (Korean-English, Urdu-English, and Chinese-English) we find substantial performance gains over the CCM model, a strong monolingual baseline. On average, across a variety of scenarios, our model achieves a 8.2 absolute increase in F-measure.

Reinforcement Learning for Mapping Instructions to Actions

S.R.K. Branavan, Harr Chen, Luke Zettlemoyer and Regina Barzilay

In this paper, we present a reinforcement learning approach for mapping natural language instructions to sequences of executable actions. We assume access to a reward function that defines the quality of the executed actions. During training, the learner repeatedly constructs action sequences for a set of documents, executes those actions, and observes the resulting reward. We use a policy gradient algorithm to estimate the parameters of a log-linear model for action selection. We apply our method to interpret instructions in two domains — Windows troubleshooting guides and game tutorials. Our results demonstrate that this method can rival supervised learning techniques while requiring few or no annotated training examples.

Learning Semantic Correspondences with Less Supervision

Percy Liang, Michael Jordan and Dan Klein

A central problem in grounded language acquisition is learning the correspondences between a rich world state and a stream of text which references that world state. To deal with the high degree of ambiguity present in this setting, we present a generative model that simultaneously segments the text into utterances and maps each utterance to a meaning representation grounded in the world state. We show that our model generalizes across three domains of increasing difficulty—Robocup sportscasting, weather forecasts (a new domain), and NFL recaps.

Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling

Daichi Mochihashi, Takeshi Yamada and Naonori Ueda

In this paper, we propose a new Bayesian model for fully unsupervised word segmentation and an efficient blocked Gibbs sampler combined with dynamic programming for inference. Our model is a nested hierarchical Pitman-Yor language model, where Pitman-Yor spelling model is embedded in the word model. We confirmed that it significantly outperforms previous reported results in both phonetic transcripts and standard datasets for Chinese and Japanese word segmentation. Our model is also considered as a way to construct an accurate word n-gram language model directly from characters of arbitrary language, without any “word” indications.

Knowing the Unseen: Estimating Vocabulary Size over Unseen Samples

Suma Bhat and Richard Sproat

Empirical studies on corpora involve making measurements of several quantities for the purpose of comparing corpora, creating language models or to make generalizations about specific linguistic phenomena in a language. Quantities such as average word length are stable across sample sizes and hence can be reliably estimated from large enough samples. However, quantities such as vocabulary size change with sample size. Thus measurements based on a given sample will need to be extrapolated to obtain their estimates over larger unseen samples. In this work, we propose a novel nonparametric estimator of vocabulary size. Our main result is to show the statistical consistency of the estimator – the first of its kind in the literature. Finally, we compare our proposal with the state of the art estimators (both parametric and nonparametric) on large standard corpora; apart from showing the favorable performance of our estimator, we also see that the classical Good-Turing estimator consistently underestimates the vocabulary size.

A Ranking Approach to Stress Prediction for Letter-to-Phoneme Conversion

Qing Dou, Shane Bergsma, Sittichai Jiampojamarn and Grzegorz Kondrak

Correct stress placement is important in text-to-speech systems, in terms of both the overall accuracy and the naturalness of pronunciation. In this paper, we formulate stress assignment as a sequence prediction problem. We represent words as sequences of substrings, and use the substrings as features in a Support Vector Machine (SVM) ranker, which is trained to rank possible stress patterns. The ranking approach facilitates inclusion of arbitrary features over both the input sequence and output stress pattern. Our system advances the current state-of-the-art, predicting primary stress in English, German, and Dutch with up to 98% word accuracy on phonemes, and 96% on letters. The system is also highly accurate in predicting secondary stress. Finally, when applied in tandem with an L2P system, it substantially reduces the word error rate when predicting both phonemes and stress.

Reducing the Annotation Effort for Letter-to-Phoneme Conversion

Kenneth Dwyer and Grzegorz Kondrak

Letter-to-phoneme (L2P) conversion is the process of producing a correct phoneme sequence for a word, given its letters. It is often desirable to reduce the quantity of training data — and hence human annotation — that is needed to train an L2P classifier for a new language. In this paper, we confront the challenge of building an accurate L2P classifier with a minimal amount of training data by combining several diverse techniques: context ordering, letter clustering, active learning, and phonetic L2P alignment. Experiments on six languages show up to 75% reduction in annotation effort.

Transliteration Alignment

Vladimir Pervouchine, Haizhou Li and Bo Lin

This paper studies transliteration alignment, its evaluation metrics and applications. We propose a new evaluation metric, alignment entropy, grounded on the information theory, to evaluate the alignment quality without the need for the gold standard reference and compare the metric with F-score. We study the use of phonological features and affinity statistics for transliteration alignment at phoneme and grapheme levels. The experiments show that better alignment consistently leads to more accurate transliteration. In transliteration modeling application, we achieve a mean reciprocal rate (MRR) of 0.773 on Xinhua personal name corpus, a significant improvement over other reported results on the same corpus. In transliteration validation application, we achieve 4.48% equal error rate on a large LDC name corpus.

Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike

Bart Jongejan and Hercules Dalianis

We propose a method to automatically train lemmatization rules that handle prefix, infix and suffix changes to generate the lemma from the full form of a word. We explain how the lemmatization rules are created and how the lemmatizer works. We trained this lemmatizer on Danish, Dutch, English, German, Greek, Icelandic, Norwegian, Polish, Slovene and Swedish full form-lemma pairs respectively. We obtained significant improvements of 24 percent for Polish, 2.3 percent for Dutch, 1.5 percent for English, 1.2 percent for German and 1.0 percent for Swedish compared to plain suffix lemmatization using a suffix-only lemmatizer. Icelandic deteriorated with 1.9 percent. We also made an observation regarding the number of produced lemmatization rules as a function of the number of training pairs.

Revisiting Pivot Language Approach for Machine Translation

Hua Wu and Haifeng Wang

This paper revisits the pivot language approach for machine translation. First, we investigate three different methods for pivot translation. Then we employ a hybrid method combining RBMT and SMT systems to fill up the data gap for pivot translation, where the source-pivot and pivot-target corpora are independent. Experimental results on spoken language translation show that this hybrid method significantly improves the translation quality, which outperforms the method using a source-target corpus of the same size. In addition, we propose a system combination approach to select better translations from those produced by various pivot translation methods. This method regards system combination as a translation evaluation problem and formalizes it with a regression learning model. Experimental results indicate that our method achieves consistent and significant improvement over individual translation outputs.

Efficient Minimum Error Rate Training and Minimum Bayes-Risk Decoding for Translation Hypergraphs and Lattices

Shankar Kumar, Wolfgang Macherey, Chris Dyer and Franz Och

Minimum Error Rate Training (MERT) and Minimum Bayes-Risk (MBR) decoding are used in most current state-of-the-art Statistical Machine Translation (SMT) systems. The algorithms were originally developed to work with N-best lists of translations, and recently extended to lattices that encode many more hypotheses than typical N-best lists. We here extend lattice-based MERT and MBR algorithms to work with hypergraphs that encode a vast number of translations produced by MT systems based on Synchronous Context Free Grammars. These algorithms are more efficient than the lattice-based versions presented earlier. We show how MERT can be employed to optimize parameters for MBR decoding. Our experiments show speedups from MERT and MBR as well as performance improvements from MBR decoding on several language pairs.

Forest-based Tree Sequence to String Translation Model

Hui Zhang, Min Zhang, Haizhou Li, Aiti Aw and Chew Lim Tan

This paper proposes a forest-based tree sequence to string translation model for syntax-based statistical machine translation, which automatically learns tree sequence to string translation rules from word-aligned source-side-parsed bilingual texts. The proposed model leverages on the strengths of both tree sequence-based and forest-based translation models. Therefore, it can not only utilize forest structure that compactly encodes exponential number of parse trees but also capture non-syntactic translation equivalences with linguistically structured information through tree sequence. This makes our model potentially more robust to parse errors and structure divergence. Experimental results on the NIST MT-2003 Chinese-English translation task show that our method statistically significantly outperforms the four baseline systems.

Active Learning for Multilingual Statistical Machine Translation

Gholamreza Haffari and Anoop Sarkar

Statistical machine translation (SMT) models require bilingual corpora for training, and these corpora are often multilingual with parallel text in multiple languages simultaneously. We introduce an *active learning* task of adding a new language to an existing multilingual set of parallel text and constructing high quality MT systems, from each language in the collection into this new target language. We show that adding a new language using active learning to the EuroParl corpus provides a significant improvement compared to a random sentence selection baseline. We also provide new highly effective sentence selection methods that improve phrase-based SMT in the multilingual and single language-pair setting.

DEPEVAL(summ): Dependency-based Evaluation for Automatic Summaries

Karolina Owczarzak

This paper presents DEPEVAL(summ), a dependency-based metric for automatic evaluation of summaries. Using a reranking parser and a Lexical-Functional Grammar (LFG) annotation, we produce a set of dependency triples for each summary. The dependency set for each candidate summary is then automatically compared against dependencies generated from model summaries. We examine a number of variations of the method, including the addition of WordNet, partial matching, or removing relation labels from the dependencies. In a test on TAC 2008 data, DEPEVAL(summ) achieves comparable or higher correlations with human judgments than the popular evaluation metrics ROUGE and Basic Elements (BE).

Summarizing Definition from Wikipedia

Shiren Ye, Tat-Seng Chua and Jie Lu

Wikipedia provides a wealth of knowledge, where the first sentence, infobox (and relevant sentences), and even the entire document of a wiki article could be considered as diverse versions of summaries (definitions) of the target topic. We explore how to generate a series of summaries with various lengths based on them. To obtain more reliable associations between sentences, we introduce wiki concepts according to the internal links in Wikipedia. In addition, we develop an extended document concept lattice model to combine wiki concepts and non-textual features such as the outline and infobox. The model can concatenate representative sentences from non-overlapping salient local topics for summary generation. We test our model based on our annotated wiki articles which topics come from TREC-QA 2004-2006 evaluations. The results show that the model is effective in summarization and definition QA.

Automatically Generating Wikipedia Articles: A Structure-Aware Approach

Christina Sauper and Regina Barzilay

In this paper, we investigate an approach for creating a comprehensive textual overview of a subject composed of information drawn from the Internet. We use the high-level structure of human-authored texts to automatically induce a domain-specific template for the topic structure of a new overview. The algorithmic innovation of our work is a method to learn topic-specific extractors for content selection jointly for the entire template. We augment the standard perceptron algorithm with a global integer linear programming formulation to optimize both local fit of information into each topic and global coherence across the entire overview. The results of our evaluation confirm the benefits of incorporating structural information into the content selection process.

Learning to Tell Tales: A Data-driven Approach to Story Generation

Neil McIntyre and Mirella Lapata

Computational story telling has sparked great interest in artificial intelligence, partly because of its relevance to educational and gaming applications. Traditionally, story generators rely on a large repository of background knowledge containing information about the story plot and its characters. This information is detailed and usually hand crafted. In this paper we propose a data-driven approach for generating short children's stories that does not require extensive manual involvement. We create an end-to-end system that realizes the various components of the generation pipeline stochastically. Our system follows a generate-and-and-rank approach where the space of multiple candidate stories is pruned by considering whether they are plausible, interesting, and coherent.

Recognizing Stances in Online Debates

Swapna Somasundaran and Janyce Wiebe

This paper presents an unsupervised opinion analysis approach for debate-side classification. In order to handle the complexities of this genre, we mine the web to learn associations that are indicators of opinion stances in debates. We combine this with discourse information and formulate the debate side classification as an Integer Linear Programming problem. Our results show that our method is substantially better than smart baselines for this task.

Co-Training for Cross-Lingual Sentiment Classification

Xiaojun Wan

The lack of Chinese sentiment corpora limits the research progress on Chinese sentiment classification. However, there are many freely available English sentiment corpora on the Web. This paper focuses on the problem of cross-lingual sentiment classification, which leverages an available English corpus for Chinese sentiment classification by using the English corpus as training data. Machine translation services are used for eliminating the language gap between the training set and test set, and English features and Chinese features are considered as two independent views of the classification problem. We propose a co-training approach to making use of unlabeled Chinese data. Experimental results show the effectiveness of the proposed approach, which can outperform the standard inductive classifiers and the transductive classifiers.

A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge

Tao Li, Yi Zhang and Vikas Sindhwani

Sentiment classification refers to the task of automatically identifying whether a given piece of text expresses positive or negative opinion towards a subject at hand. The proliferation of user-generated web content such as blogs, discussion forums and online review sites has made it possible to perform large-scale mining of public opinion. Sentiment modeling is thus becoming a critical component of market intelligence and social media technologies that aim to tap into the collective wisdom of crowds. In this paper, we consider the problem of learning high-quality sentiment models with minimal manual supervision. We propose a novel approach to learn from lexical prior knowledge in the form of domain-independent sentiment-laden terms, in conjunction with unlabeled data. Our model is based on a constrained non-negative tri-factorization of the term-document matrix which can be implemented using simple update rules. Extensive experimental studies demonstrate the effectiveness of our approach on a variety of real-world sentiment prediction tasks.

Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis

Jungi Kim, Jin-Ji Li and Jong-Hyeok Lee

This paper describes an approach to utilizing term weights for sentiment analysis tasks and shows how various term weighting schemes improve the performance of sentiment analysis systems. Previously, sentiment analysis was mostly studied under data-driven and lexicon-based frameworks. Such work generally exploits textual features for fact-based analysis tasks or lexical indicators from a sentiment lexicon. We propose to model term weighting into a sentiment analysis system utilizing collection statistics, contextual and topic-related characteristics as well as opinion-related properties. Experiments carried out on various datasets show that our approach effectively improves previous methods.

Compiling a Massive, Multilingual Dictionary via Probabilistic Inference

Mausam, Stephen Soderland, Oren Etzioni, Daniel Weld, Michael Skinner and Jeff Bilmes

Can we automatically compose a large set of Wiktionaries and translation dictionaries to yield a massive, multilingual dictionary whose coverage is substantially greater than that of any of its constituent dictionaries? The composition of multiple translation dictionaries leads to a transitive inference problem: if word A translates to word B which in turn translates to word C, what is the probability that C is a translation of A? The paper introduces a novel algorithm that solves this problem for 10,000,000 words in more than 1,000 languages. The algorithm yields PANDICTIONARY, a novel multilingual dictionary. PANDICTIONARY contains more than four times as many translations than in the largest Wiktionary at precision 0.90 and over 200,000,000 pairwise translations in over 200,000 language pairs at precision 0.8.

A Metric-based Framework for Automatic Taxonomy Induction

Hui Yang and Jamie Callan

This paper presents a novel metric-based framework for the task of automatic taxonomy induction. The framework incrementally clusters terms based on ontology metric, a score indicating semantic distance; and transforms the task into a multi-criterion optimization based on minimization of taxonomy structures and modeling of term abstractness. It combines the strengths of both lexico-syntactic patterns and clustering through incorporating heterogeneous features. The flexible design of the framework allows a further study on which features are the best for the task under various conditions. The experiments not only show that our system achieves higher F1-measure than other state-of-the-art systems, but also reveal the interaction between features and various types of relations, as well as the interaction between features and term abstractness.

Learning with Annotation Noise

Eyal Beigman and Beata Beigman Klebanov

It is usually assumed that the kind of noise existing in annotated data is random classification noise. Yet there is evidence that differences between annotators are not always random attention slips but could result from different biases towards the classification categories, at least for the harder-to-decide cases. Under an annotation generation model that takes this into account, there is a hazard that some of the training instances are actually hard cases with unreliable annotations. We show that these are relatively unproblematic for an algorithm operating under the 0-1 loss model, whereas for the commonly used voted perceptron algorithm, hard training cases could result in incorrect prediction on the uncontroversial cases at test time.

Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both?

Paola Merlo and Lonneke van der Plas

Semantic role labels are the representation of the grammatically relevant aspects of a sentence meaning. Capturing the nature and the number of semantic roles in a sentence is therefore fundamental to correctly describing the interface between grammar and meaning. In this paper, we compare two annotation schemes, PropBank and VerbNet, in a task-independent, general way, analysing how well they fare in capturing the linguistic generalisations that are known to hold for semantic role labels, and consequently how well they grammaticalise aspects of meaning. We show that VerbNet is more verb-specific and better able to generalise to new semantic role instances, while PropBank better captures some of the structural constraints among roles. We conclude that these two resources should be used together, as they are complementary.

Robust Machine Translation Evaluation with Entailment Features

Sebastian Pado, Michel Galley, Dan Jurafsky and Christopher D. Manning

Existing evaluation metrics for machine translation lack crucial robustness: their correlations with human quality judgments vary considerably across languages and genres. We believe that the main reason is their inability to properly capture meaning: A good translation candidate means the same thing as the reference translation, regardless of formulation. We propose a metric that evaluates MT output based on a rich set of features motivated by textual entailment, such as lexical-semantic (in-)compatibility and argument structure overlap. We compare this metric against a combination metric of four state-of-the-art scores (BLEU, NIST, TER, and METEOR) in two different settings. The combination metric outperforms the individual scores, but is bested by the entailment-based metric. Combining the entailment and traditional features yields further improvements.

The Contribution of Linguistic Features to Automatic Machine Translation Evaluation

Enrique Amigó, Jesús Giménez, Julio Gonzalo and Felisa Verdejo

A number of approaches to Automatic MT Evaluation based on deep linguistic knowledge have been suggested. However, n-gram based metrics are still today the dominant approach. The main reason is that the advantages of employing deeper linguistic information have not been clarified yet. In this work, we propose a novel approach for meta-evaluation of MT evaluation metrics, since correlation cofficient against human judges do not reveal details about the advantages and disadvantages of particular metrics. We then use this approach to investigate the benefits of introducing linguistic features into evaluation metrics. Overall, our experiments show that (i) both lexical and linguistic metrics present complementary advantages and (ii) combining both kinds of metrics yields the most robust meta-evaluation performance.

A Syntax-Driven Bracketing Model for Phrase-Based Translation

Deyi Xiong, Min Zhang, Aiti Aw and Haizhou Li

Syntactic analysis influences the way in which the source sentence is translated. Previous efforts add syntactic constraints to phrase-based translation by directly rewarding/punishing a hypothesis whenever it matches/violates source-side constituents. We present a new model that automatically learns syntactic constraints, including but not limited to constituent matching/violation, from training corpus. The model brackets a source phrase as to whether it satisfies the learnt syntactic constraints. The bracketed phrases are then translated as a whole unit by the decoder. Experimental results and analysis show that the new model outperforms other previous methods and achieves a substantial improvement over the baseline which is not syntactically informed.

Topological Ordering of Function Words in Hierarchical Phrase-based Translation

Hendra Setiawan, Min-Yen Kan, Haizhou Li and Philip Resnik

Hierarchical phrase-based models are attractive because they provide a consistent framework within which to characterize both local and long-distance reorderings, but they also make it difficult to distinguish many implausible reorderings from those that are linguistically plausible. Rather than appealing to annotation-driven syntactic modeling, we address this problem by observing the influential role of function words in determining syntactic structure, and introducing soft constraints on function word relationships as part of a standard log-linear hierarchical phrase-based model. Experimentation on Chinese-English and Arabic-English translation demonstrates that the approach yields significant gains in performance.

Phrase-Based Statistical Machine Translation as a Traveling Salesman Problem

Mikhail Zaslavskiy, Marc Dymetman and Nicola Cancedda

An efficient decoding algorithm is a crucial element of any statistical machine translation system. Some researchers have noted certain similarities between SMT decoding and the famous Traveling Salesman Problem, in particular (Knight, 1999) has shown that any TSP instance may be mapped to a sub-case of a word-based SMT model, demonstrating NP-hardness of the decoding task. In this paper, we focus on the reverse mapping, showing that any phrase-based SMT decoding problem can be directly reformulated as a TSP. The transformation is very natural, deepens our understanding of the decoding problem, and allows direct use of any of the powerful existing TSP solvers for SMT decoding. We test our approach on three datasets, where TSP-based decoders are compared to the popular beam-search algorithm. In all cases, our method provides competitive or better performance.

Concise Integer Linear Programming Formulations for Dependency Parsing

Andre Martins, Noah A. Smith and Eric Xing

We formulate the problem of non-projective dependency parsing as a polynomial-sized integer linear program. Our formulation is able to handle non-local output features in an efficient manner; not only is it compatible with prior knowledge encoded as hard constraints, it can also learn "soft" constraints from data. In particular, our model is able to capture correlations among neighboring arcs (siblings and grandparents), word valency, and can learn to favor nearly-projective parses according to what is observed in the data. The model parameters are learned in a max-margin framework by employing a linear programming relaxation. We evaluate the performance of our parser on data in several natural languages, achieving improvements over existing state-of-the-art methods.

Non-Projective Dependency Parsing in Expected Linear Time

Joakim Nivre

We present a novel transition system for dependency parsing, which constructs arcs only between adjacent words but can parse arbitrary non-projective trees by swapping the order of words in the input. Adding the swapping operation changes the time complexity for deterministic parsing from linear to quadratic in the worst case, but empirical estimates based on treebank data show that the expected running time is in fact linear for the range of data attested in the corpora. Evaluation of the parsing system on data from five languages shows state-of-the-art accuracy, with especially good results for the labeled exact match score.

Semi-supervised Learning of Dependency Parsers using Generalized Expectation Criteria

Gregory Druck, Gideon Mann and Andrew McCallum

In this paper, we propose a novel method for semi-supervised learning of non-projective log-linear dependency parsers using directly expressed linguistic prior knowledge (e.g. a noun's parent is often a verb). Model parameters are estimated using a generalized expectation (GE) objective function that penalizes the mismatch between model predictions and linguistic expectation constraints. In a comparison with two prominent "unsupervised" learning methods that require indirect biasing toward the correct syntactic structure, we show that GE can attain better accuracy with as few as 20 intuitive constraints. We also present positive experimental results on longer sentences in multiple languages.

Dependency Grammar Induction via Bitext Projection Constraints

Kuzman Ganchev, Jennifer Gillenwater and Ben Taskar

Broad-coverage annotated treebanks necessary to train parsers do not exist for many resource-poor languages. The wide availability of parallel text and accurate parsers in English has opened up the possibility of grammar induction through partial transfer across bitext. We consider generative and discriminative models for dependency grammar induction that use word-level alignments and a source language parser (English) to constrain the space of possible target trees. Unlike previous approaches, our framework does not require full projected parses, allowing partial, approximate transfer through linear expectation constraints on the space of distributions over trees. We consider several types of constraints that range from generic dependency conservation to language-specific annotation rules for auxiliary verb analysis. We evaluate our approach on Bulgarian and Spanish CoNLL shared task data and show that we consistently outperform unsupervised methods and can outperform supervised learning for limited training data.

Cross-Domain Dependency Parsing Using a Deep Linguistic Grammar

Yi Zhang and Rui Wang

Pure statistical parsing systems achieves high in-domain accuracy but performs poorly out-domain. In this paper, we propose two different approaches to produce syntactic dependency structures using a large-scale hand-crafted HPSG grammar. The dependency backbone of an HPSG analysis is used to provide general linguistic insights which, when combined with state-of-the-art statistical dependency parsing models, achieves performance improvements on out-domain tests.

A Chinese-English Organization Name Translation System Using Heuristic Web Mining and Asymmetric Alignment

Fan Yang, Jun Zhao and Kang Liu

In this paper, we propose a novel system for translating organization names from Chinese to English with the assistance of web resources. Firstly, we adopt a chunking-based segmentation method to improve the segmentation of Chinese organization name which is plagued by the OOV problem. Then a heuristic query construction method is employed to construct an efficient query which can be used to search the bilingual Web pages containing translation equivalents. Finally, we align the Chinese organization name with English sentences using the asymmetric alignment method to find the best English fragment as the translation equivalent. The experimental results show that the proposed method can outperform the baseline statistical machine translation system by 30.44%.

Reducing Semantic Drift with Bagging and Distributional Similarity

Tara McIntosh and James R. Curran

Iterative bootstrapping algorithms are typically compared using a single set of hand-picked seeds. However, we demonstrate that performance varies greatly depending on these seeds, and favourable seeds for one algorithm can perform very poorly with others, making comparisons unreliable. We exploit this wide variation with bagging, sampling from automatically extracted seeds to reduce semantic drift. However, semantic drift still occurs in later iterations. We propose an integrated distributional similarity filter to identify and censor potential semantic drifts, ensuring over 10% higher precision when extracting large semantic lexicons.

Jointly Identifying Temporal Relations with Markov Logic

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara and Yuji Matsumoto

Recent work on temporal relation identification has focused on three types of relations between events: temporal relations between an event and a time expression, between a pair of events and between an event and the document creation time. These types of relations have mostly been identified in isolation by event pairwise comparison. However, this approach neglects logical constraints between temporal relations of different types that we believe to be helpful. We therefore propose a Markov Logic model that jointly identifies relations of all three relation types simultaneously. By evaluating our model on the TempEval data we show that this approach leads to about 2% higher accuracy for all three types of relations —and to the best results for the task when compared to those of other machine learning based systems.

Profile Based Cross-Document Coreference Using Kernelized Soft Relational Clustering

Jian Huang, Sarah M. Taylor, Jonathan L. Smith, Konstantinos A. Fotiadis and C. Lee Giles

Coreferencing entities across documents in a large corpus enables advanced document understanding tasks such as question answering. This paper presents a novel cross document coreference approach that leverages the profiles of entities which are constructed by using information extraction tools and reconciled by using a within-document coreference module. We propose to match the profiles by using a learned ensemble distance function comprised of a suite of similarity specialists. We develop a kernelized soft relational clustering algorithm that makes use of the learned distance function to partition the entities into fuzzy sets of identities. We compare the kernelized clustering method with a popular fuzzy relation clustering algorithm (FRC) and show 5% improvement in coreference performance. Evaluation of our proposed methods on a large benchmark disambiguation collection shows that they compare favorably with the top runs in the SemEval evaluation.

Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task

Kristen Parton, Kathleen R. McKeown, Bob Coyne, Mona T. Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gokhan Tur, Wei Xu and Sibel Yaman

Cross-lingual tasks are especially difficult due to the compounding effect of errors in language processing and errors in machine translation (MT). In this paper, we present an error analysis of a new cross-lingual task: the 5W task, a sentence-level understanding task which seeks to return the English 5W's (Who, What, When, Where and Why) corresponding to a Chinese sentence. We analyze systems that we developed, identifying specific prob-lems in language processing and MT that cause errors. The best cross-lingual 5W sys-tem was still 19% worse than the best mono-lingual 5W system, which shows that MT significantly degrades sentence-level under-standing. Neither source-language nor target-language analysis was able to circumvent problems in MT, although each approach had advantages relative to the other. A detailed error analysis across multiple systems sug-gests directions for future research on the problem.

Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition

Jong-Hoon Oh, Kiyotaka Uchimoto and Kentaro Torisawa

This paper proposes a novel framework called *bilingual co-training* for a large-scale, accurate acquisition method for *monolingual* semantic knowledge. In this framework, we combine the independent processes of monolingual semantic-knowledge acquisition for two languages using bilingual resources to boost performance. We apply this framework to large-scale hyponymy-relation acquisition from Wikipedia. Experimental results show that our approach improved the F-measure by 3.6–10.3%. We also show that bilingual co-training enables us to build classifiers for two languages in tandem with the same combined amount of data as required for training a single classifier in isolation while achieving superior performance.

Automatic Set Instance Extraction using the Web

Richard C. Wang and William W. Cohen

An important and well-studied problem is the production of semantic lexicons from a large corpus. In this paper, we present a system named ASIA (Automatic Set Instance Acquirer), which takes in the name of a semantic class as input (e.g., “car makers”) and automatically outputs its instances (e.g., “ford”, “nissan”, “toyota”). ASIA is based on recent advances in web-based set expansion - the problem of finding all instances of a set given a small number of “seed” instances. This approach effectively exploits web resources and can be easily adapted to different languages. In brief, we use language-dependent hyponym patterns to find a noisy set of initial seeds, and then use a state-of-the-art language-independent set expansion system to expand these seeds. The proposed approach matches or outperforms prior systems on several English-language benchmarks. It also shows excellent performance on three dozen additional benchmark problems from English, Chinese and Japanese, thus demonstrating language-independence.

Extracting Lexical Reference Rules from Wikipedia

Eyal Shnarch, Libby Barak and Ido Dagan

This paper describes the extraction from Wikipedia of lexical reference rules, identifying references to term meanings triggered by other terms. We present extraction methods geared to cover the broad range of the lexical reference relation and analyze them extensively. Most extraction methods yield high precision levels, and our rule-base is shown to perform better than other automatically constructed baselines in a couple of lexical expansion and matching tasks. Our rule-base yields comparable performance to WordNet while providing largely complementary information.

Employing Topic Models for Pattern-based Semantic Class Discovery

Huibin Zhang, Mingjie Zhu, Shuming Shi and Ji-Rong Wen

A semantic class is a collection of items (words or phrases) which have semantically peer or sibling relationship. This paper studies the employment of topic models to automatically construct semantic classes, taking as the source data a collection of raw semantic classes (RASCs), which were extracted by applying predefined patterns to web pages. The primary requirement (and challenge) here is dealing with multi-membership: An item may belong to multiple semantic classes; and we need to discover as many as possible the different semantic classes the item belongs to. To adopt topic models, we treat RASCs as “documents”, items as “words”, and the final semantic classes as “topics”. Appropriate pre-processing and postprocessing are performed to improve results quality, to reduce computation cost, and to tackle the fixed-k constraint of a typical topic model. Experiments conducted on 40 million web pages show that our approach could yield better results than alternative approaches.

Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition

Dipanjan Das and Noah A. Smith

We present a novel approach to deciding whether two sentences hold a paraphrase relationship. We employ a generative model that generates a paraphrase of a given sentence, and we use probabilistic inference to reason about whether two sentences share the paraphrase relationship. The model cleanly incorporates both syntax and lexical semantics using quasi-synchronous dependency grammars (Smith and Eisner, 2006). Furthermore, using a product of experts (Hinton, 2002), we combine the model with a complementary logistic regression model based on state-of-the-art lexical overlap features. We evaluate our models on the task of distinguishing true paraphrase pairs from false ones on a standard corpus, giving competitive state-of-the-art performance.

Full Paper Presentations – Tuesday, 4 August

Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty

Yoshimasa Tsuruoka, Jun'ichi Tsujii and Sophia Ananiadou

Stochastic gradient descent (SGD) uses approximate gradients estimated from subsets of the training data and updates the parameters in an online fashion. This learning framework is attractive because it often requires much less training time in practice than batch training algorithms. However, L1-regularization, which is becoming popular in natural language processing because of its ability to produce compact models, cannot be efficiently applied in SGD training, due to the large dimensions of feature vectors and the fluctuations of approximate gradients. We present a simple method to solve these problems by penalizing the weights according to cumulative values for L1 penalty. We evaluate the effectiveness of our method in three applications: text chunking, named entity recognition, and part-of-speech tagging. Experimental results demonstrate that our method can produce compact and accurate models much more quickly than a state-of-the-art quasi-Newton method for L1-regularized loglinear models.

A global model for joint lemmatization and part-of-speech prediction

Kristina Toutanova and Colin Cherry

We present a global joint model for lemmatization and part-of-speech prediction. Using only morphological lexicons and unlabeled data, we learn a partially-supervised part-of-speech tagger and a lemmatizer which are combined using features on a dynamically linked dependency structure of words. We evaluate our model on English, Bulgarian, Czech, and Slovene, and demonstrate substantial improvements over both a direct transduction approach to lemmatization and a pipelined approach, which predicts part-of-speech tags before lemmatization.

Distributional Representations for Handling Sparsity in Supervised Sequence-Labeling

Fei Huang and Alexander Yates

Supervised sequence-labeling systems in natural language processing often suffer from data sparsity because they use word types as features in their prediction tasks. Consequently, they have difficulty estimating parameters for types which appear in the test set, but seldom (or never) appear in the training set. We demonstrate that distributional representations of word types, trained on unannotated text, can be used to improve performance on rare words. We incorporate aspects of these representations into the feature space of our sequence-labeling systems. In an experiment on a standard chunking dataset, our best technique improves a chunker from 0.76 F1 to 0.86 F1 on chunks beginning with rare words. On the same dataset, it improves our part-of-speech tagger from 74% to 80% accuracy on rare words. Furthermore, our system improves significantly over a baseline system when applied to text from a different domain, and it reduces the sample complexity of sequence labeling.

Minimized Models for Unsupervised Part-of-Speech Tagging

Sujith Ravi and Kevin Knight

We describe a novel method for the task of unsupervised POS tagging with a dictionary, one that uses integer programming to explicitly search for the smallest model that explains the data, and then uses EM to set parameter values. We evaluate our method on a standard test corpus using different standard tagsets (a 45-tagset as well as a smaller 17-tagset), and show that our approach performs better than existing state-of-the-art systems in both settings.

An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa and Hitoshi Isahara

In this paper, we present a discriminative word-character hybrid model for joint Chinese word segmentation and POS tagging. Our word-character hybrid model offers high performance since it can handle both known and unknown words. We describe our strategies that yield good balance for learning the characteristics of known and unknown words and propose an error-driven policy that delivers such balance by acquiring examples of unknown words from particular errors in a training corpus. We describe an efficient framework for training our model based on the Margin Infused Relaxed Algorithm (MIRA), evaluate our approach on the Penn Chinese Treebank, and show that it achieves superior performance compared to the state-of-the-art approaches reported in the literature.

Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging – A Case Study

Wenbin Jiang, Liang Huang and Qun Liu

Manually annotated corpora are valuable but scarce resources, yet for many annotation tasks such as treebanking and sequence labeling there exist multiple corpora with different and incompatible annotation guidelines or standards. This seems to be a great waste of human efforts, and it would be nice to automatically adapt one annotation standard to another. We present a simple yet effective strategy that transfers knowledge from a differently annotated corpus to the corpus with desired annotation. We test the efficacy of this method in the context of Chinese word segmentation and part-of-speech tagging, where no segmentation and POS tagging standards are widely accepted due to the lack of morphology in Chinese. Experiments show that adaptation from the much larger Peoples Daily corpus to the smaller but more popular Penn Chinese Treebank results in significant improvements in both segmentation and tagging accuracies (with error reductions of 30.2% and 14%, respectively), which in turn helps improve Chinese parsing accuracy.

Linefeed Insertion into Japanese Spoken Monologue for Captioning

Tomohiro Ohno, Masaki Murata and Shigeki Matsubara

To support the real-time understanding of spoken monologue such as lectures and commentaries, the development of a captioning system is required. In monologues, since a sentence tends to be long, each sentence is often displayed in multi lines on one screen, it is necessary to insert linefeeds into a text so that the text becomes easy to read. This paper proposes a technique for inserting linefeeds into a Japanese spoken monologue text as an elemental technique to generate the readable captions. Our method appropriately inserts linefeeds into a sentence by machine learning, based on the information such as dependencies, clause boundaries, pauses and line length. An experiment using Japanese speech data has shown the effectiveness of our technique.

Semi-supervised Learning for Automatic Prosodic Event Detection Using Co-training Algorithm

Je Hun Jeon and Yang Liu

Most of previous approaches to automatic prosodic event detection are based on supervised learning, relying on the availability of a corpus that is annotated with the prosodic labels of interest in order to train the classification models. However, creating such resources is an expensive and time-consuming task. In this paper, we exploit semi-supervised learning with the co-training algorithm for automatic detection of coarse level representation of prosodic events such as pitch accents, intonational phrase boundaries, and break indices. We propose a confidence-based method to assign labels to unlabeled data and demonstrate improved results using this method compared to the widely used agreement-based method. In addition, we examine various informative sample selection methods. In our experiments on the Boston University radio news corpus, using only a small amount of the labeled data as the initial training set, our proposed labeling method combined with most confidence sample selection can effectively use unlabeled data to improve performance and finally reach performance closer to that of the supervised method using all the training data.

Summarizing multiple spoken documents: finding evidence from untranscribed audio

Xiaodan Zhu, Gerald Penn and Frank Rudzicz

This paper presents a model for summarizing multiple spoken documents over untranscribed audio stream. Without assuming the availability of transcripts, the model modifies a recently proposed unsupervised algorithm to detect re-occurring acoustic patterns in speech and uses them to estimate similarities between utterances, which are in turn used to identify salient utterances and remove redundancies. This model is of interest due to its independence from spoken language transcripts, an error-prone and resource-intensive process, its ability to integrate multiple sources of information on the same topic, and its novel use of acoustic patterns that extends previous work on low-level prosodic feature detection. We compare the performance of this model with that achieved using manual and automatic transcripts, and find that this new approach is roughly equivalent to having access to ASR transcripts with word error rates in the 32-36% range without actually having to do the ASR, plus it better handles utterances with out-of-vocabulary words.

(Note: Short Paper abstracts not included)

Improving Tree-to-Tree Translation with Packed Forests

Yang Liu, Yajuan Lü and Qun Liu

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

Fast Consensus Decoding over Translation Forests

John DeNero, David Chiang and Kevin Knight

The minimum Bayes risk (MBR) decoding objective improves BLEU scores for machine translation output relative to the standard Viterbi objective of maximizing model score. However, MBR targeting BLEU is prohibitively slow to optimize over k-best lists for large k. In this paper, we introduce and analyze an alternative to MBR that is equally effective at improving performance, yet is asymptotically faster — running 80 times faster than MBR in experiments with 1000-best lists. Furthermore, our fast decoding procedure can select output sentences based on distributions over entire forests of translations, in addition to k-best lists. We evaluate our procedure on translation forests from two large-scale, state-of-the-art hierarchical machine translation systems. Our forest-based decoding objective consistently outperforms k-best list MBR, giving improvements of up to 1.0 BLEU.

Joint Decoding with Multiple Translation Models

Yang Liu, Haitao Mi, Yang Feng and Qun Liu

Current SMT systems usually decode with single translation models and cannot benefit from the strengths of other models in decoding phase. We instead propose joint decoding, a method that combines multiple translation models in one decoder. Our joint decoder draws connections among multiple models by integrating the translation hypergraphs they produce individually. Therefore, one model can share translations and even derivations with other models. Comparable to the state-of-the-art system combination technique, joint decoding achieves an absolute improvement of 1.5 BLEU points over individual decoding.

Collaborative Decoding: Partial Hypothesis Re-ranking Using Translation Consensus between Decoders

Mu Li, Nan Duan, Dongdong Zhang, Chi-Ho Li and Ming Zhou

This paper presents collaborative decoding (co-decoding), a new method to improve machine translation accuracy by leveraging translation consensus between multiple machine translation decoders. Different from system combination and MBR decoding, which post-process the n-best lists or word lattice of machine translation decoders, in our method multiple machine translation decoders collaborate by exchanging partial translation results. Using an iterative decoding approach, n-gram agreement statistics between translations of multiple decoders are employed to re-rank both full and partial hypothesis explored in decoding. Experimental results on data sets for NIST Chinese-to-English machine translation task show that the co-decoding method can bring significant improvements to all baseline decoders, and the outputs from co-decoding can be used to further improve the result of system combination.

Variational Decoding for Statistical Machine Translation

Zhifei Li, Jason Eisner and Sanjeev Khudanpur

Statistical models in machine translation exhibit spurious ambiguity. That is, the probability of an output string is split among many distinct derivations (e.g., trees or segmentations). In principle, the goodness of a string is measured by the total probability of its many derivations. However, finding the best string (e.g., during decoding) is then computationally intractable. Therefore, most systems use a simple Viterbi approximation that measures the goodness of a string using only its most probable derivation. Instead, we develop a variational approximation, which considers all the derivations but still allows tractable decoding. Our particular variational distributions are parameterized as n-gram models. We also analytically show that interpolating these n-gram models for different n is similar to minimum-risk decoding for BLEU (Tromble et al., 2008). Experiments show that our approach improves the state of the art.

Unsupervised Learning of Narrative Schemas and their Participants

Nathanael Chambers and Dan Jurafsky

We describe an unsupervised system for learning narrative schemas, coherent sequences or sets of events (arrested(Police, Suspect)), convicted(Judge, Suspect)) whose arguments are filled with participant semantic roles defined over words (Judge = {judge, jury, court}, Police = {police, agent, authorities}). Unlike most previous work in event structure or semantic role learning, our system does not use supervised techniques, hand-built knowledge, or predefined classes of events or roles. Our unsupervised learning algorithm uses coreferring arguments in chains of verbs to learn both rich narrative event structure and argument roles. By jointly addressing both tasks, we improve on previous results in narrative/frame learning and induce rich frame-specific semantic roles.

Learning a Compositional Semantic Parser using an Existing Syntactic Parser

Ruifang Ge and Raymond Mooney

We present a new approach to learning a semantic parser (a system that maps natural language sentences into logical form). Unlike previous methods, it exploits an existing syntactic parser to produce disambiguated parse trees that drive the compositional semantic interpretation. The resulting system produces improved results on standard corpora on natural language interfaces for database querying and simulated robot control.

Latent Variable Models of Concept-Attribute Attachment

Joseph Reisinger and Marius Pasca

This paper presents a set of Bayesian methods for automatically extending the WordNet ontology with new concepts and annotating existing concepts with generic property fields, or attributes. We base our approach on Latent Dirichlet Allocation and evaluate along three dimensions: (1) the precision of the ranked lists of attributes, (2) the quality of the attribute assignments to WordNet concepts, and (3) the specificity of the attributes at each concept. In all cases we find that the principled LDA-based approaches outperform previously proposed heuristic methods.

The Chinese Aspect Generation Based on Aspect Selection Functions

Guowen Yang and John Bateman

This paper describes our system for generating Chinese aspect expressions. In the system, the semantics of different aspects is characterized by specific temporal and conceptual features. The two kinds of features build semantic applicability conditions for each individual aspect. The semantic applicability conditions of a specific aspect are theoretically represented by the Aspect Selection Function (ASF)of the aspect. The ASF takes relevant time points and concepts as its parameters and “calculates” the truth value of the semantic applicability conditions of the aspect in order to make a corresponding aspect selection in language generation. Our system has been implemented as a grammar for the KPML multilingual generator which is equipped with a large systemic grammar and all the technical components required in generation. Fourteen primary simple aspects, and twenty-six complex aspects are organized into a hierarchical system network. The generation is realized by evaluating implemented inquiries which formally define the ASFs, traversing the grammatical network, and making aspect selections. Each path through the network from the root to an end node corresponds to a specific language expression. In the present research the application of the ASFs provides a formal way to represent semantic applicability conditions of the aspects; the grammatical network built on the basis of systemic functional grammar systematically organizes and distinguishes semantic properties of different aspects. The computational implementation verifies both grammatical organization and semantic descriptions of the Chinese aspects.

Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation

Kai-min K. Chang, Vladimir L. Cherkassky, Tom M. Mitchell and Marcel Adam Just

Recent advances in functional Magnetic Resonance Imaging (fMRI) offer a significant new approach to studying semantic representations in humans by making it possible to directly observe brain activity while people comprehend words and sentences. In this study, we investigate how humans comprehend adjective-noun phrases (e.g. strong dog) while their neural activity is recorded. Classification analysis shows that the distributed pattern of neural activity contains sufficient signal to decode differences among phrases. Furthermore, vector-based semantic models can explain a significant portion of systematic variance in the observed neural activity. Multiplicative composition models of the two-word phrase outperform additive models, consistent with the assumption that people use adjectives to modify the meaning of the noun, rather than conjoining the meaning of the adjective and noun.

Capturing Salience with a Trainable Cache Model for Zero-anaphora Resolution

Ryu Iida, Kentaro Inui and Yuji Matsumoto

Many researchers have recently explored machine learning-based methods for anaphora resolution. In those approaches, the resolution is carried out by searching for an antecedent from a set of candidates which appears in the preceding context of a given anaphor. However, it involves a serious drawback because the search space tends to expand greatly when the text is longer. This problem is addressed in theory-oriented rule-based approaches based on Centering Theory (Grosz et al. 1995) in the sense that their models choose an antecedent from a limited number of salient candidates at each point of unfolding discourse. In those approaches, updating discourse status is performed by hand-crafted rules based on human insights. However, such rule-based approaches need to be sophisticated to adapt to real texts. In order to solve this problem, we propose to employ a machine learning paradigm to optimize the process of updating discourse status empirically. We recast the task of updating the discourse status as the problem of ranking discourse entities by adopting the notion of caching originally introduced by Walker (1996). More specifically, we choose salient candidates for each sentence from the set of candidates appearing in that sentence and the candidates which are already in the cache. Using this mechanism, the computational cost of the zero-anaphora resolution process is reduced by searching only the set of salient candidates. Our empirical evaluation on Japanese zero-anaphora resolution shows that our learning-based cache model drastically reduces the search space while preserving the accuracy.

Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art

Veselin Stoyanov, Nathan Gilbert, Claire Cardie and Ellen Riloff

We aim to shed light on the state-of-the-art in NP coreference resolution by teasing apart the differences in the MUC and ACE task definitions, the assumptions made in evaluation methodologies, and inherent differences in text corpora. First, we examine three subproblems that play a role in coreference resolution: named entity recognition, anaphoricity determination, and coreference element detection. We measure the impact of each subproblem on coreference resolution, and confirm that certain assumptions regarding these subproblems in the evaluation methodology can dramatically simplify the overall task. Second, we measure the performance of a state-of-the-art coreference resolver on several classes of anaphora and use these results to develop a quantitative measure for estimating coreference resolution performance on new data sets.

A Novel Discourse Parser Based on Support Vector Machine Classification

David duVerle and Helmut Prendinger

This paper introduces a new algorithm to parse discourse within the framework of Rhetorical Structure Theory (RST). Our method is based on recent advances in the field of statistical machine learning (multivariate capabilities of Support Vector Machines) and a rich feature space. RST offers a formal framework for hierarchical text organization with strong applications in discourse analysis and text generation. We demonstrate automated annotation of a text with RST hierarchically organised relations, with results comparable to those achieved by specially trained human annotators. Using a rich set of shallow lexical, syntactic and structural features from the input text, our parser achieves, in linear time, 73.9% of professional annotators' agreement F-score. The parser is 5% to 12% more accurate than current state-of-the-art parsers.

Genre distinctions for discourse in the Penn TreeBank

Bonnie Webber

Articles in the Penn TreeBank were identified as being reviews, summaries, letters to the editor, news reportage, corrections, wit and short verse, or quarterly profit reports. All but the latter three were then characterised in terms of features manually annotated in the Penn Discourse TreeBank — discourse connectives and their senses. Summaries turned out to display very different discourse features than the other three genres. Letters also appeared to have some different features. The two main findings involve (1) differences between genres in the senses associated with intra-sentential discourse connectives, inter-sentential discourse connectives and inter-sentential discourse relations that are not lexically marked; and (2) differences within all four genres between the senses of discourse relations not lexically marked and those that are marked. The first finding means that genre should be made a factor in automated sense labelling of non-lexically marked discourse relations. The second means that lexically marked relations provide a poor model for automated sense labelling of relations that are not lexically marked.

Automatic sense prediction for implicit discourse relations in text

Emily Pitler, Annie Louis and Ani Nenkova

We present a series of experiments on automatically identifying the sense of implicit discourse relations, i.e. relations that are not marked with a discourse connective such as but or because. We work with a corpus of implicit relations present in newspaper text and report results on a test set that is representative of the naturally occurring distribution of senses. We use several linguistically informed features, including polarity tags, Levin verb classes, length of verb phrases, modality, context, and lexical features. In addition, we revisit past approaches using lexical pairs from unannotated text as features, explain some of their shortcomings and propose modifications. Our best combination of features outperforms the baseline from data intensive approaches by 4% for comparison and 16% for contingency.

(Note: Student Research Workshop abstracts not included)

(Note: Short Paper abstracts not included)

A Framework of Feature Selection Methods for Text Categorization

Shoushan Li, Rui Xia, Chengqing Zong and Chu-Ren Huang

In text categorization, feature selection (FS) is a strategy that aims at making text classifiers more efficient and accurate. However, when dealing with a new task, it is still difficult to quickly select a suitable one from various FS methods provided by many previous studies. In this paper, we propose a theoretic framework of FS methods based on two basic measurements: frequency measurement and ratio measurement. Then six popular FS methods are in detail discussed under this framework. Moreover, with the guidance of our theoretical analysis, we propose a novel method called weighed frequency and odds (WFO) that combines the two measurements with trained weights. The experimental results on data sets from both topic-based and sentiment classification tasks show that this new method is robust across different tasks and numbers of selected features.

Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification

Sajib Dasgupta and Vincent Ng

Supervised polarity classification systems are typically domain-specific. Building these systems involves collecting and annotating a large amount of data for each domain, making the construction of polarity classifiers prohibitively expensive. A potential solution to this problem is to build unsupervised polarity classification systems. However, unsupervised classification of sentiment is difficult: reviews are mostly ambiguous, as reviewers normally discuss both the positive and negative aspects of a product. We need a sophisticated learning system that can handle the ambiguous reviews effectively. To address this problem, we propose a mostly-unsupervised approach to sentiment classification that involves a novel combination of spectral learning, active learning and bootstrapping. Experimental results on five sentiment classification datasets show that we can automatically identify a large number of unambiguous reviews with high precision using our approach and employ them to improve a polarity classifier while minimizing human intervention.

Modeling Latent Biographic Attributes in Conversational Genres

Nikesh Garera and David Yarowsky

This paper presents and evaluates several original techniques for the latent classification of biographic attributes such as gender, age and native language, in diverse genres (conversation transcripts, email) and languages (Arabic, English). First, we present a novel partner-sensitive model for extracting biographic attributes in conversations, given the differences in lexical usage and discourse style such as observed between same-gender and mixed-gender conversations. Then, we explore a rich variety of novel sociolinguistic and discourse-based features, including mean utterance length, passive/active usage, percentage domination of the conversation, speaking rate and filler word usage. Cumulatively up to 20% error reduction is achieved relative to the standard Boulis and Ostendorf (2005) algorithm for classifying individual conversations, and accuracy for gender detection on the Switchboard Corpus (aggregate) and Gulf Arabic Corpus exceeds 95%.

A Graph-based Semi-Supervised Learning for Question-Answering

Asli Celikyilmaz, Marcus Thint and Zhiheng Huang

We present a graph-based semi-supervised learning for the question-answering (QA) task for ranking candidate sentences. Using textual entailment analysis, we obtain entailment scores between a natural language question posed by the user and the candidate sentences returned from search engine. The textual entailment between two sentences is assessed via features representing high-level attributes of the entailment problem such as sentence structure matching, question-type named-entity matching based on a question-classifier, etc. We implement a semi-supervised learning (SSL) approach to demonstrate that utilization of more unlabeled data points can improve the answer-ranking task of QA. We create a graph for labeled and unlabeled data using match-scores of textual entailment features as similarity weights between data points. We apply a summarization method on the graph to make the computations feasible on large datasets. With a new representation of graph-based SSL on QA datasets using only a handful of features, and under limited amounts of labeled data, we show improvement in generalization performance over state-of-the-art QA models.

Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding

Delphine Bernhard and Iryna Gurevych

Monolingual translation probabilities have recently been introduced in retrieval models to solve the lexical gap problem. They can be obtained by training statistical translation models on parallel monolingual corpora, such as question-answer pairs, where answers act as the “source” language and questions as the “target” language. In this paper, we propose to use as a parallel training dataset the definitions and glosses provided for the same term by different lexical semantic resources. We compare monolingual translation models built from lexical semantic resources with two other kinds of datasets: manually-tagged question reformulations and question-answer pairs. We also show that the monolingual translation probabilities obtained (i) are comparable to traditional semantic relatedness measures and (ii) significantly improve the results over the query likelihood and the vector-space model for answer finding.

Answering Opinion Questions with Random Walks on Graphs

Fangtao Li, Yang Tang, Minlie Huang and Xiaoyan Zhu

Opinion Question Answering (*Opinion QA*), which aims to find the authors’ sentimental opinions on a specific target, is more challenging than traditional fact-based question answering problems. To extract the opinion oriented answers, we need to consider both topic relevance and opinion sentiment issues. Current solutions to this problem are mostly ad-hoc combinations of question topic information and opinion information. In this paper, we propose an *Opinion PageRank* model and an *Opinion HITS* model to fully explore the information from different relations among questions and answers, answers and answers, and topics and opinions. By fully exploiting these relations, the experiment results show that our proposed algorithms outperform several state of the art baselines on benchmark data set. A gain of over 10% in F scores is achieved as compared to many other systems.

What lies beneath: Semantic and syntactic analysis of manually reconstructed spontaneous speech

Erin Fitzgerald, Frederick Jelinek and Robert Frank

Spontaneously produced speech text often includes disfluencies which make it difficult to analyze underlying structure. Successful reconstruction of this text would transform these errorful utterances into fluent strings and offer an alternate mechanism for analysis. Our investigation of naturally-occurring spontaneous speaker errors aligned to corrected text with manual semantico-syntactic analysis yields new insight into the syntactic and structural semantic differences between spoken and reconstructed language.

Discriminative Lexicon Adaptation for Improved Character Accuracy - A New Direction in Chinese Language Modeling

Yi-Cheng Pan, Lin-Shan Lee and Sadaoki Furui

While OOV is always a problem for most languages in ASR, in the Chinese case the problem can be avoided by utilizing character n-grams and moderate performances can be obtained. However, character ngram has its own limitation and proper addition of new words can increase the ASR performance. Here we propose a discriminative lexicon adaptation approach for improved character accuracy, which not only adds new words but also deletes some words from the current lexicon. Different from other lexicon adaptation approaches, we consider the acoustic features and make our lexicon adaptation criterion consistent with that in the decoding process. The proposed approach not only improves the ASR character accuracy but also significantly enhances the performance of a character-based spoken document retrieval system.

Improving Automatic Speech Recognition for Lectures through Transformation-based Rules Learned from Minimal Data

Cosmin Munteanu, Gerald Penn and Xiaodan Zhu

We demonstrate that transformation-based learning can be used to correct noisy speech recognition transcripts in the lecture domain with an average word error rate reduction of 12.9%. Our method is distinguished from earlier related work by its robustness to small amounts of training data, and its resulting efficiency, in spite of its use of true word error rate computations as a rule scoring function.

(Note: Short Paper abstracts not included)

Full Paper Presentations – Wednesday, 5 August

Quadratic-Time Dependency Parsing for Machine Translation

Michel Galley and Christopher D. Manning

Efficiency is a prime concern in syntactic MT decoding, yet significant developments in statistical parsing with respect to asymptotic efficiency haven't yet been explored in MT. Recently, McDonald et al. (2005) formalized dependency parsing as a maximum spanning tree (MST) problem, which can be solved in quadratic time relative to the length of the sentence. They show that MST parsing is almost as accurate as cubic-time dependency parsing in the case of English, and that it is more accurate with free word order languages. This paper applies MST parsing to MT, and describes how it can be integrated into a phrase-based decoder to compute dependency language model scores. Our results show that augmenting a state-of-the-art phrase-based system with this dependency language model leads to significant improvements in TER (0.92%) and BLEU (0.45%) scores on five NIST Chinese-English evaluation test sets.

A Gibbs Sampler for Phrasal Synchronous Grammar Induction

Phil Blunsom, Trevor Cohn, Chris Dyer and Miles Osborne

In this paper we present a phrasal synchronous grammar model of translational equivalence. Unlike previous approaches, we do not resort to heuristics or constraints from word-alignment models, but instead directly induce a synchronous grammar from parallel sentence-aligned corpora. We use a hierarchical Bayesian prior to bias towards compact grammars with small translation units. Inference is performed using a novel Gibbs sampler over synchronous derivations. This sampler side-steps intractability issues with inference over derivation forests - required by previous models - which has time complexity of at least $O(n^6)$. Instead each sampling iteration is highly efficient and mixes well, allowing the model to be applied to larger machine translation corpora than previous approaches.

Source-Language Entailment Modeling for Translating Unknown Terms

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman and Idan Szpektor

This paper addresses the task of handling unknown terms in SMT. We propose using source-language monolingual models and resources to paraphrase the source text prior to translation. We further present a conceptual extension to prior work by allowing translations of entailed texts rather than paraphrases only. A method for performing this process efficiently is presented and applied to some 2500 sentences with unknown terms. Our experiments show that the proposed approach substantially increases the number of properly translated texts.

Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh and Pushpak Bhattacharyya

We report in this paper our work on accurately generating case markers and suffixes in English-to-Hindi SMT. Hindi is a relatively free word-order language, and makes use of a comparatively richer set of case markers and morphological suffixes for correct meaning representation. From our experience of large-scale English-Hindi MT, we are convinced that fluency and fidelity in the Hindi output get an order of magnitude facelift if accurate case markers and suffixes are produced. Now, the moot question is: *what entity on the English side encodes the information contained in case markers and suffixes on the Hindi side?* Our studies of correspondences in the two languages show that case markers and suffixes in Hindi are predominantly determined by the combination of suffixes and semantic relations on the English side. We, therefore, augment the aligned corpus of the two languages, with the correspondence of English suffixes and semantic relations with Hindi suffixes and case markers. Our results on 400 test sentences, translated using an SMT system trained on around 13000 parallel sentences, show that *suffix + semantic relation → case marker/suffix* is a very useful translation factor, in the sense of making a significant difference to output quality as indicated by subjective evaluation as well as BLEU scores.

Dependency Based Chinese Sentence Realization

Wei He, Haifeng Wang, Yuqing Guo and Ting Liu

This paper describes log-linear models for a general-purpose sentence realizer based on dependency structures. Unlike traditional realiz-ers using grammar rules, our method realizes sentences by linearizing dependency relations directly in two steps. First, the relative order between head and each dependent is deter-mined by their dependency relation. Then the best linearizations compatible with the relative order are selected by log-linear models. The log-linear models incorporate three types of feature functions, including dependency rela-tions, surface words and headwords. Our ap-proach to sentence realization provides sim-plicity, efficiency and competitive accuracy. Trained on 8,975 dependency structures of a Chinese Dependency Treebank, the realizer achieves a BLEU score of 0.8874.

Incorporating Information Status into Generation Ranking

Aoife Cahill and Arndt Riester

We investigate the influence of information status (IS) on constituent order in German, and integrate our findings into a log-linear surface realisation ranking model. We show that the distribution of pairs of IS categories is strongly asymmetric. Moreover, each category is correlated with morphosyntactic features, which can be automatically detected. We build a log-linear model that incorporates these asymmetries for ranking German string realisations from input LFG F-structures. We show that it achieves a statistically significantly higher BLEU score than the baseline system without these features.

A Syntax-Free Approach to Japanese Sentence Compression

Tsutomu Hirao, Jun Suzuki and Hideki Isozaki

Conventional sentence compression methods require a syntactic parser to compress a sentence without changing its meaning. However, reference compressions made by humans do not always retain the syntactic structures of the original sentences. Moreover, for the goal of on-demand sentence compression, the time spent in the parsing stage is not negligible. As an alternative to syntactic parsing, we propose a novel term weighting technique based on the positional information within the original sentence and a novel language model that combines statistics from the original sentence and a general corpus. Experiments that involve both human subjective evaluations and automatic evaluations show that our method outperforms Hori's method, a state-of-the-art conventional technique. Moreover, because our method does not use a syntactic parser, it is 4.3 times faster than Hori's method.

Application-driven Statistical Paraphrase Generation

Shiqi Zhao, Xiang Lan, Ting Liu and Sheng Li

Paraphrase generation (PG) is important in plenty of NLP applications. However, the research of PG is far from enough. In this paper, we propose a novel method for statistical paraphrase generation (SPG), which can (1) achieve various applications based on a uniform statistical model, and (2) naturally combine multiple resources to enhance the PG performance. In our experiments, we use the proposed method to generate paraphrases for three different applications. The results show that the method can be easily transformed from one application to another and generate valuable and interesting paraphrases.

Semi-Supervised Cause Identification from Aviation Safety Reports

Isaac Persing and Vincent Ng

We introduce cause identification, a new problem involving classification of incident reports in the aviation domain. Specifically, given a set of pre-defined causes, a cause identification system seeks to identify all and only those causes that can explain why the aviation incident described in a given report occurred. The difficulty of cause identification stems in part from the fact that it is a multi-class, multi-label categorization task, and in part from the skewness of the class distributions and the scarcity of annotated reports. To improve the performance of a cause identification system for the minority classes, we present a bootstrapping algorithm that automatically augments a training set by learning from a small amount of labeled data and a large amount of unlabeled data. Experimental results show that our algorithm yields a relative error reduction of 6.3% in F-measure for the minority classes in comparison to a baseline that learns solely from the labeled data.

SMS based Interface for FAQ Retrieval

Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkatesan T. Chakaravarthy and L. Venkata Subramaniam

Short Messaging Service (SMS) is popularly used to provide information access to people on the move. This has resulted in the growth of SMS based Question Answering (QA) services. However automatically handling SMS questions poses significant challenges due to the inherent noise in SMS questions. In this work we present an automatic FAQ-based question answering system for SMS users. We handle the noise in a SMS query by formulating the query similarity over FAQ questions as a combinatorial search problem. The search space consists of combinations of all possible dictionary variations of tokens in the noisy query. We present an efficient search algorithm that does not require any training data or SMS normalization and can handle semantic variations in question formulation. We demonstrate the effectiveness of our approach on two reallife datasets.

Semantic Tagging of Web Search Queries

Mehdi Manshadi and Xiao Li

We give a novel approach to parse web search queries for the purpose of automatic tagging of the queries. Inspired by GPSG's ID/LP rules, we define a set of probabilistic free-order context-free rules, which generates bags (multi-sets) of words. Using this new type of rule in combination with the tradi-tional probabilistic phrase structure rules, we define a hybrid grammar, which treats every search query as a bag of chunks (phrases). A hybrid probabilistic parser is used to parse the queries. In order to taking contextual informa-tion into account, a discriminative classifier is used on top of the parser, re-ranking the n-best parse trees generated by the parser. Experi-ments show that our approach outperforms our basic model, which is based on Condi-tional Random Fields.

Mining Bilingual Data from the Web with Adaptively Learnt Patterns

Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu and Qingsheng Zhu

Mining bilingual data (including bilingual sentences and terms) from the Web can benefit many NLP applications, such as machine translation and cross language information retrieval. In this paper, based on the observation that bilingual data in many web pages appear collectively following similar patterns, an adaptive pattern-based bilingual data mining method is proposed. Specifically, given a web page, the method contains four steps: 1) preprocessing: parse the web page into a DOM tree and segment the inner text of each node into snippets; 2) seed mining: identify potential translation pairs (seeds) using a word based alignment model which takes both translation and transliteration into consideration; 3) pattern learning: learn generalized patterns with the identified seeds; 4) pattern based mining: extract all bilingual data in the page using the learned patterns. Our experiments on Chinese web pages produced more than 7.5 million pairs of bilingual sentences and more than 5 million pairs of bilingual terms, both with over 80% accuracy.

Comparing Objective and Subjective Measures of Usability in a Human-Robot Dialogue System

Mary Ellen Foster, Manuel Giuliani and Alois Knoll

We present a human-robot dialogue system that enables a robot to work together with a human user to build wooden construction toys. We then describe a study in which naive subjects interacted with this system under a range of conditions and then completed a user-satisfaction questionnaire. The results of this study provide a wide range of subjective and objective measures of the quality of the interactions. To assess which aspects of the interaction had the greatest impact on the users' opinions of the system, we used a method based on the PARADISE evaluation framework (Walker et al., 1997) to derive a performance function from our data. The major contributors to user satisfaction were the number of repetition requests (which had a negative effect on satisfaction), the dialogue length, and the users' recall of the system instructions (both of which contributed positively).

Setting Up User Action Probabilities in User Simulations for Dialog System Development

Hua Ai and Diane Litman

User simulations are shown to be useful in spoken dialog system development. Since most current user simulations deploy probability models to mimic human user behaviors, how to set up user action probabilities in these models is a key problem to solve. One generally used approach is to estimate these probabilities from human user data. However, when building a new dialog system, usually no data or only a small amount of data is available. In this study, we compare estimating user probabilities from a small user data set versus handcrafting the probabilities. We discuss the pros and cons of both solutions for different dialog system development tasks.

Dialogue Segmentation with Large Numbers of Volunteer Internet Annotators

T. Daniel Midgley

This paper shows the results of an experiment in dialogue segmentation. In this experiment, segmentation was done on a level of analysis similar to adjacency pairs. The method of annotation was somewhat novel: volunteers were invited to participate over the Web, and their responses were aggregated using a simple voting method. Though volunteers received a minimum of training, the aggregated responses of the group showed very high agreement with expert opinion. The group, as a unit, performed at the top of the list of annotators, and in many cases performed as well as or better than the best annotator.

Robust Approach to Abbreviating Terms: A Discriminative Latent Variable Model with Global Information

Xu Sun, Naoaki Okazaki and Jun'ichi Tsujii

This paper describes a robust approach for abbreviating terms. First, to incorporate non-local information into abbreviation generation tasks, we present both implicit and explicit solutions: the latent variable model, or alternatively, the label encoding approach with global information. Although the two approaches compete with one another, we demonstrate that they are also complementary. By combining the two, experiments revealed that our abbreviation generator achieved the best results in both Chinese and English. Moreover, we directly apply our generator to perform a very different task from tradition, abbreviation recognition. Experiments showed that our model worked robustly, and outperformed five of six state-of-the-art abbreviation recognizers.

A non-contiguous Tree Sequence Alignment-based Model for Statistical Machine Translation

Jun Sun, Min Zhang and Chew Lim Tan

The tree sequence based translation model allows the violation of syntactic boundaries in a rule to capture non-syntactic phrases, where a tree sequence is a contiguous sequence of sub-trees. This paper goes further to present a translation model based on non-contiguous tree sequence alignment, where a non-contiguous tree sequence is a sequence of sub-trees and gaps. Compared with the contiguous tree sequence-based model, the proposed model can well handle non-contiguous phrases with any large gaps by means of non-contiguous tree sequence alignment. An algorithm targeting the non-contiguous constituent decoding is also proposed. Experimental results on the NIST MT-05 Chinese-English translation task show that the proposed model statistically significantly outperforms the baseline systems.

Better Word Alignments with Supervised ITG Models

Aria Haghighi, John Blitzer, John DeNero and Dan Klein

We investigate supervised word alignment methods that exploit inversion transduction grammar (ITG) constraints. We show that 1-to-1 ITG models outperform models over all 1-to-1 matchings when using both heuristic and learned parameters. Additionally, we show that allowing many-to-one alignment blocks further improves prediction accuracy. We consider discriminative methods for maximizing both margin and conditional likelihood,describing how both procedures can be adapted to achieve state-of-the-art alignment results. In total, these improvements result in the best published Chinese-English alignment performance and yield a downstream translation improvement of 1.1 BLEU over GIZA++ alignments.

Confidence Measure for Word Alignment

Fei Huang

In this paper we present a confidence measure for word alignment based on the posterior probability of alignment links. We introduce sentence alignment confidence measure and alignment link confidence measure. Based on these measures, we improve the alignment quality by selecting high confidence sentence alignments and alignment links from multiple word alignments of the same sentence pair. Additionally, we remove low confidence alignment links from the word alignment of a bilingual training corpus, which increases the alignment F-score, improves Chinese-English and Arabic-English translation quality and significantly reduces the phrase translation table size.

A Comparative Study of Hypothesis Alignment and its Improvement for Machine Translation System Combination

Boxing Chen, Min Zhang, Haizhou Li and Aiti Aw

Recently confusion network decoding shows the best performance in combining outputs from multiple machine translation (MT) systems. However, overcoming different word orders presented in multiple MT systems during hypothesis alignment still remains the biggest challenge to confusion network-based MT system combination. In this paper, we compare four commonly used word alignment methods, namely GIZA++, TER, CLA and IHMM, for hypothesis alignment. Then we propose a method to build the confusion network from intersection word alignment, which utilizes both direct and inverse word alignment between the backbone and hypothesis to improve the reliability of hypothesis alignment. Experimental results demonstrate that the intersection word alignment yields consistent performance improvement for all four word alignment methods on both Chinese-to-English spoken and written language tasks.

Incremental HMM Alignment for MT System Combination

Chi-Ho Li, Xiaodong He, Yupeng Liu and Ning Xi

Inspired by the incremental TER alignment, we re-designed the Indirect HMM (IHMM) alignment, which is one of the best alignment methods for conventional MT system combination, in incremental manner. One crucial problem of incremental alignment is to align a hypothesis to a confusion network (CN). Our incremental IHMM alignment is implemented in three different ways: 1) treat CN spans as HMM states and define state transition as distortion over covered n-grams between two spans; 2) treat CN spans as HMM states and define state transition as distortion over words in component translations; and 3) use a consensus decoding algorithm over multiple IHMMs, each of which corresponds to a component translation in the CN. All these three approaches of incremental alignment based on IHMM are shown to be superior to both incremental TER alignment and conventional IHMM alignment in the setting of the Chinese-to-English track of the 2008 NIST Open MT evaluation.

K-Best A* Parsing

Adam Pauls and Dan Klein

A* parsing makes 1-best search efficient by suppressing unlikely 1-best items. Existing k-best extraction methods can efficiently search for top derivations, but only after an exhaustive 1-best pass. We present a unified algorithm for k-best A* parsing which preserves the efficiency of k-best extraction while giving the speed-ups of A* parsing methods. Our algorithm produces optimal k-best parses under the same conditions required for optimality in a standard A* parser. Empirically, optimal k-best lists can be extracted significantly faster than with other approaches, over a range of grammar types.

Coordinate Structure Analysis with Global Structural Constraints and Alignment-Based Local Features

Kazuo Hara, Masashi Shimbo, Hideharu Okuma and Yuji Matsumoto

We propose a hybrid approach to coordinate structure analysis that combines a simple grammar to ensure consistent global structure of coordinations in a sentence, and features based on sequence alignment to capture local symmetry of conjuncts. The weight of the alignment-based features, which in turn determines the score of coordinate structures, is optimized by perceptron training on a given corpus. A bottom-up chart parsing algorithm efficiently finds the best scoring structure, taking both nested or non-overlapping flat coordinations into account. We demonstrate that our approach outperforms existing parsers in coordination scope detection on the Genia corpus.

Learning Context-Dependent Mappings from Sentences to Logical Form

Luke Zettlemoyer and Michael Collins

We consider the problem of learning context-dependent mappings from sentences to logical form. The training examples are sequences of sentences annotated with lambda-calculus meaning representations. We develop an algorithm that maintains explicit, lambda-calculus representations of salient discourse entities and uses a context-dependent analysis pipeline to recover logical forms. The method uses a hidden-variable variant of the perception algorithm to learn a linear model used to select the best analysis. Experiments on context-dependent utterances from the ATIS corpus show that the method recovers fully correct logical forms with 83.7% accuracy.

An Optimal-Time Binarization Algorithm for Linear Context-Free Rewriting Systems with Fan-Out Two

Carlos Gómez-Rodríguez and Giorgio Satta

Linear context-free rewriting systems (LCFRSs) are grammar formalisms with the capability of modeling discontinuous constituents. Many applications use LCFRSs where the fan-out (a measure of the discontinuity of phrases) is not allowed to be greater than 2. We present an efficient algorithm for transforming LCFRS with fan-out at most 2 into a binary form, whenever this is possible. This results in asymptotical run-time improvement for known parsing algorithms for this class.

A Polynomial-Time Parsing Algorithm for TT-MCTAG

Laura Kallmeyer and Giorgio Satta

This paper investigates the class of Tree-Tuple MCTAG with Shared Nodes, TT-MCTAG for short, an extension of Tree Adjoining Grammars that has been proposed for natural language processing, in particular for dealing with discontinuities and word order variation in languages such as German. It has been shown that the universal recognition problem for this formalism is NP-complete, but so far it was not known whether the class of languages generated by TT-MCTAG is included in PTIME. We provide a positive answer to this question, using a new characterization of TT-MCTAG and exploiting it in the development of a novel polynomial parsing algorithm for this class.

Distant supervision for relation extraction without labeled data

Mike Mintz, Steven Bills, Rion Snow and Dan Jurafsky

Modern models of relation extraction for tasks like ACE are based on supervised learning of relations from small hand-labeled corpora. We investigate an alternative paradigm that does not require labeled corpora, avoiding the domain dependence of ACE-style algorithms, and allowing the use of corpora of any size. Our experiments use Freebase, a large semantic database of several thousand relations, to provide distant supervision. For each pair of entities that appears in some Freebase relation, we find all sentences containing those entities in a large unlabeled corpus and extract textual features to train a relation classifier. Our algorithm combines the advantages of supervised IE (combining 400,000 noisy pattern features in a probabilistic classifier) and unsupervised IE (extracting large numbers of relations from large corpora of any domain). Our model is able to extract 10,000 instances of 102 relations at a precision of 67.6%. We also analyze feature performance, showing that syntactic parse features are particularly helpful for relations that are ambiguous or lexically distant in their expression.

Multi-Task Transfer Learning for Weakly-Supervised Relation Extraction

Jing Jiang

Creating labeled training data for relation extraction is expensive. In this paper, we study relation extraction in a special weakly-supervised setting when we have only a few seed instances of the target relation type we want to extract but we also have a large amount of labeled instances of other relation types. Observing that different relation types can share certain common structures, we propose to use a multi-task learning method coupled with human guidance to address this weakly-supervised relation extraction problem. The proposed framework models the commonality among different relation types through a shared weight vector, enables knowledge learned from the auxiliary relation types to be transferred to the target relation type, and allows easy control of the tradeoff between precision and recall. Empirical evaluation on the ACE 2004 data set shows that the proposed method substantially improves over two baseline methods.

Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web

Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang and Mitsuru Ishizuka

This paper presents an unsupervised relation extraction method for discovering and enhancing relations in which a specified concept in Wikipedia participates. Using respective characteristics of Wikipedia articles and Web corpus, we develop a clustering approach based on combinations of patterns: dependency patterns from dependency analysis of texts in Wikipedia, and surface patterns generated from highly redundant information related to the Web. Evaluations of the proposed approach on two different domains demonstrate the superiority of the pattern combination over existing approaches. Fundamentally, our method demonstrates how deep linguistic patterns contribute complementarily with Web surface patterns to the generation of various relations.

Phrase Clustering for Discriminative Learning

Dekang Lin and Xiaoyun Wu

We present a simple and scalable algorithm for clustering tens of millions of phrases and use the resulting clusters as features in discriminative classifiers. To demonstrate the power and generality of this approach, we apply the method in two very different applications: named entity recognition and query classification. Our results show that phrase clusters offer significant improvements over word clusters. Our NER system achieves the best current result on the widely used CoNLL benchmark. Our query classifier is on par with the best system in KDDCUP 2005 without resorting to labor intensive knowledge engineering efforts.

Semi-Supervised Active Learning for Sequence Labeling

Katrin Tomanek and Udo Hahn

While Active Learning (AL) has already been shown to dramatically reduce the annotation effort for many sequence labeling tasks compared to random selection, AL remains agnostic about the internal structure of the selected sequences typically, sentences). We here propose a semi-supervised approach to AL for sequence labeling where only very uncertain subsequences are presented to a human for annotation, while all other subsequences within these selected sequences are automatically labeled. For the task of named entity recognition, our experiments reveal that this approach reduces the annotation effort in terms of manually labeled tokens by up to 60% compared to the standard, fully supervised AL scheme.

Word or Phrase? Learning Which Unit to Stress for Information Retrieval

Young-In Song, Jung-Tae Lee and Hae-Chang Rim

The use of phrases in retrieval models has been proven to be helpful in the literature, but no particular research addresses the problem of discriminating phrases that are likely to degrade the retrieval performance from the ones that do not. In this paper, we present a retrieval framework that utilizes both words and phrases flexibly, followed by a general learning-to-rank method for learning the potential contribution of a phrase in retrieval. We also present useful features that reflect the compositionality and discriminative power of a phrase and its constituent words for optimizing the weights of phrase use in phrase-based retrieval models. Experimental results on the TREC collections show that our proposed method is effective.

A Generative Blog Post Retrieval Model that Uses Query Expansion based on External Collections

Wouter Weerkamp, Krisztian Balog and Maarten de Rijke

User generated content is characterized by short, noisy documents, with many spelling errors and unexpected language usage. To bridge the vocabulary gap between the user's information need and documents in a specific user generated content environment, the blogosphere, we apply a form of query expansion, i.e., adding and reweighing query terms. Since the blogosphere is noisy, query expansion on the collection itself is rarely effective but external, edited collections are more suitable. We propose a generative model for expanding queries using external collections in which dependencies between queries, documents, and expansion documents are explicitly modeled. Different instantiations of our model are discussed and make different (in)dependence assumptions. Results using two external collections (news and Wikipedia) show that external expansion for retrieval of user generated content is effective; besides, conditioning the external collection on the query is very beneficial, and making candidate expansion terms dependent on just the document seems sufficient.

Language Identification of Search Engine Queries

Hakan Ceylan and Yookyung Kim

We consider the language identification problem for search engine queries. First, we propose a method to automatically generate a data set, which uses click-through logs of the Yahoo! Search Engine to derive the language of a query indirectly from the language of the documents clicked by the users. Next, we use this data set to train two decision tree classifiers; one that only uses linguistic features and is aimed for textual language identification, and one that additionally uses a non-linguistic feature, and is geared towards the identification of the language intended by the users of the search engine. Our results show that our method produces a highly reliable data set very efficiently, and our decision tree classifier outperforms some of the best methods that have been proposed for the task of written language identification on the domain of search engine queries.

Exploiting Bilingual Information to Improve Web Search

Wei Gao, John Blitzer, Ming Zhou and Kam-Fai Wong

Web search quality can vary widely across languages, even for the same information need. We propose to exploit this variation in quality by learning a ranking function on bilingual queries: queries that appear in query logs for two languages but represent equivalent search interests. For a given bilingual query, along with corresponding monolingual query log and monolingual ranking, we generate a ranking on pairs of documents, one from each language. Then we learn a linear ranking function which exploits bilingual features on pairs of documents, as well as standard monolingual features. Finally, we show how to reconstruct monolingual ranking from a learned bilingual ranking. Using publicly available Chinese and English query logs, we demonstrate for both languages that our ranking technique exploiting bilingual data leads to significant improvements over a state-of-the-art monolingual ranking algorithm.

12

EMNLP Abstracts

Oral Presentations – Thursday, 6 August

Unsupervised Semantic Parsing

Hoifung Poon and Pedro Domingos

We present the first unsupervised approach to the problem of learning a semantic parser, using Markov logic. Our USP system transforms dependency trees into quasi-logical forms, recursively induces lambda forms from these, and clusters them to abstract away syntactic variations of the same meaning. The MAP semantic parse of a sentence is obtained by recursively assigning its parts to lambda-form clusters and composing them. We evaluate our approach by using it to extract a knowledge base from biomedical abstracts and answer questions. USP substantially outperforms TextRunner, DIRT and an informed baseline on both precision and recall on this task.

Graph Alignment for Semi-Supervised Semantic Role Labeling

Hagen Fürstenau and Mirella Lapata

Unknown lexical items present a major obstacle to the development of broad-coverage semantic role labeling systems. We address this problem with a semi-supervised learning approach which acquires training instances for unseen verbs from an unlabeled corpus. Our method relies on the hypothesis that unknown lexical items will be structurally and semantically similar to known items for which annotations are available. Accordingly, we represent known and unknown sentences as graphs, formalize the search for the most similar verb as a graph alignment problem and solve the optimization using integer linear programming. Experimental results show that role labeling performance for unknown lexical items improves with training data produced automatically by our method.

Semi-supervised Semantic Role Labeling Using the Latent Words Language Model

Koen Deschacht and Marie-Francine Moens

Semantic Role Labeling (SRL) has proved to be a valuable tool for performing automatic analysis of natural language texts. Currently however, most systems rely on a large training set, which is manually annotated, an effort that needs to be repeated whenever different languages or a different set of semantic roles is used in a certain application. A possible solution for this problem is semi-supervised learning, where a small set of training examples is automatically expanded using unlabeled texts. We present the Latent Words Language Model, which is a language model that learns word similarities from unlabeled texts. We use these similarities for different semi-supervised SRL methods as additional features or to automatically expand a small training set. We evaluate the methods on the PropBank dataset and find that for small training sizes our best performing system achieves an error reduction of 33.27% F1-measure compared to a state-of-the-art supervised baseline.

Semantic Dependency Parsing of NomBank and PropBank: An Efficient Integrated Approach via a Large-scale Feature Selection

Hai Zhao, Wenliang Chen and Chunyu Kit

We present an integrated dependency-based semantic role labeling system for English from both NomBank and PropBank. By introducing assistant argument labels and considering much more feature templates, two optimal feature template sets are obtained through an effective feature selection procedure and help construct a high performance single SRL system. From the evaluations on the date set of CoNLL-2008 shared task, the performance of our system is quite close to the state of the art. As to our knowledge, this is the first integrated SRL system that achieves a competitive performance against previous pipeline systems.

First- and Second-Order Expectation Semirings with Applications to Minimum-Risk Training on Translation Forests

Zhifei Li and Jason Eisner

Many statistical translation models can be regarded as weighted logical deduction. Under this paradigm, we use weights from the expectation semiring (Eisner, 2002), to compute first-order statistics (e.g., the expected hypothesis length or feature counts) over packed forests of translations (lattices or hypergraphs). We then introduce a novel second-order expectation semiring, which computes second-order statistics (e.g., the variance of the hypothesis length or the gradient of entropy). This second-order semiring is essential for many interesting training paradigms such as minimum risk, deterministic annealing, active learning, and semi-supervised learning, where gradient descent optimization requires computing the gradient of entropy or risk. We use these semirings in an open-source machine translation toolkit, Joshua, enabling minimum-risk training for a benefit of up to 1.0 BLEU point.

Feasibility of Human-in-the-loop Minimum Error Rate Training

Omar F. Zaidan and Chris Callison-Burch

Minimum error rate training (MERT) involves choosing parameter values for a machine translation (MT) system that maximize performance on a tuning set as measured by an automatic evaluation metric, such as BLEU. The method is best when the system will eventually be evaluated using the same metric, but in reality, most MT evaluations have a human-based component. Although performing MERT with a human-based metric seems like a daunting task, we describe a new metric, RYPT, which takes human judgments into account, but only requires human input to build a database that can be reused over and over again, hence eliminating the need for human input at tuning time. In this investigative study, we analyze the diversity (or lack thereof) of the candidates produced during MERT, we describe how this redundancy can be used to our advantage, and show that RYPT is a better predictor of translation quality than BLEU.

Cube Pruning as Heuristic Search

Mark Hopkins and Greg Langmead

Cube pruning is a fast inexact method for generating the items of a beam decoder. In this paper, we show that cube pruning is essentially equivalent to A* search on a specific search space with specific heuristics. We use this insight to develop faster and exact variants of cube pruning.

Effective Use of Linguistic and Contextual Information for Statistical Machine Translation

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas and Ralph Weischedel

Current methods of using lexical features in machine translation have difficulty in scaling up to realistic MT tasks due to a prohibitively large number of parameters involved. In this paper, we propose methods of using new linguistic and contextual features that do not suffer from this problem and apply them in a state-of-the-art hierarchical MT system. The features used in this work are non-terminal labels, non-terminal length distribution, source string context and source dependency LM scores. The effectiveness of our techniques is demonstrated by significant improvements over a strong baseline. On Arabic-to-English translation, improvements in lower-cased BLEU are 2.0 on NIST MT06 and 1.7 on MT08 newswire data on decoding output. On Chinese-to-English translation, the improvements are 1.0 on MT06 and 0.8 on MT08 newswire data.

Active Learning by Labeling Features

Gregory Druck, Burr Settles and Andrew McCallum

Methods that learn from prior information about input features such as generalized expectation (GE) have been used to train accurate models with very little effort. In this paper, we propose an active learning approach in which the machine solicits "labels" on features rather than instances. In both simulated and real user experiments on two sequence labeling tasks we show that our active learning method outperforms passive learning with features as well as traditional active learning with instances. Preliminary experiments suggest that novel interfaces which intelligently solicit labels on multiple features facilitate more efficient annotation.

Efficient kernels for sentence pair classification

Fabio Massimo Zanzotto and Lorenzo Dell'Arciprete

In this paper, we propose a novel class of graphs, the tripartite directed acyclic graphs (tDAGs), to model first-order rule feature spaces for sentence pair classification. We introduce a novel algorithm for computing the similarity in first-order rewrite rule feature spaces. Our algorithm is extremely efficient and, as it computes the similarity of instances that can be represented in explicit feature spaces, it is a valid kernel function.

Graphical Models over Multiple Strings

Markus Dreyer and Jason Eisner

We study graphical modeling in the case of string-valued random variables. Whereas a weighted finite-state transducer can model the probabilistic relationship between two strings, we are interested in building up joint models of three or more strings. This is needed for inflectional paradigms in morphology, cognate modeling or language reconstruction, and multiple-string alignment. We propose a Markov Random Field in which each factor (potential function) is a weighted finite-state machine, typically a transducer that evaluates the relationship between just two of the strings. The full joint distribution is then a product of these factors. Though decoding is actually undecidable in general, we can still do efficient joint inference using approximate belief propagation; the necessary computations and messages are all finite-state. We demonstrate the methods by jointly predicting morphological forms.

Reverse Engineering of Tree Kernel Feature Spaces

Daniele Pighin and Alessandro Moschitti

We present a framework to extract the most important features (tree fragments) from a Tree Kernel (TK) space according to their importance in the target kernel-based machine, e.g. Support Vector Machines (SVMs). In particular, our mining algorithm selects the most relevant features based on SVM estimated weights and uses this information to automatically infer an explicit representation of the input data. The explicit features (a) improve our knowledge on the target problem domain and (b) make large-scale learning practical, improving training and test time, while yielding accuracy in line with traditional TK classifiers. Experiments on semantic role labeling and question classification illustrate the above claims

A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora

Makoto Miwa, Rune Stre, Yusuke Miyao and Jun'ichi Tsujii

Because of the importance of protein-protein interaction (PPI) extraction from text, many corpora have been proposed with slightly differing definitions of proteins and PPI. Since no single corpus is large enough to saturate a machine learning system, it is necessary to learn from multiple different corpora. In this paper, we propose a solution to this challenge. We designed a rich feature vector, and we applied a support vector machine modified for corpus weighting (SVM-CW) to complete the task of multiple corpora PPI extraction. The rich feature vector, made from multiple useful kernels, is used to express the important information for PPI extraction, and the system with our feature vector was shown to be both faster and more accurate than the original kernel-based system, even when using just a single corpus. SVM-CW learns from one corpus, while using other corpora for support. SVM-CW is simple, but it is more effective than other methods that have been successfully applied to other NLP tasks earlier. With the feature vector and SVM-CW, our system achieved the best performance among all state-of-the-art PPI extraction systems reported so far.

Generalized Expectation Criteria for Bootstrapping Extractors using Record-Text Alignment

Kedar Bellare and Andrew McCallum

Traditionally, machine learning approaches for information extraction require human annotated data that can be costly and time-consuming to produce. However, in many cases, there already exists a database (DB) with schema related to the desired output, and records related to the expected input text. We present a conditional random field (CRF) that aligns tokens of a given DB record and its realization in text. The CRF model is trained using only the available DB and unlabeled text with generalized expectation criteria. An annotation of the text induced from inferred alignments is used to train an information extractor. We evaluate our method on a citation extraction task in which alignments between DBLP database records and citation texts are used to train an extractor. Experimental results demonstrate an error reduction of 35% over a previous state-of-the-art method that uses heuristic alignments.

Nested Named Entity Recognition

Jenny Rose Finkel and Christopher D. Manning

Many named entities contain other named entities inside them. Despite this fact, the field of named entity recognition has almost entirely ignored nested named entity recognition, but due to technological, rather than ideological reasons. In this paper, we present a new technique for recognizing nested named entities, by using a discriminative constituency parser. To train the model, we transform each sentence into a tree, with constituents for each named entity (and no other syntactic structure). We present results on both newspaper and biomedical corpora which contain nested named entities. In three out of four sets of experiments, our model outperforms a standard semi-CRF on the more traditional top-level entities. At the same time, we improve the overall F-score by up to 30% over the flat model, which is unable to recover any nested entities.

A Unified Model of Phrasal and Sentential Evidence for Information Extraction

Siddharth Patwardhan and Ellen Riloff

Information Extraction (IE) systems that extract role fillers for events typically look at the local context surrounding a phrase when deciding whether to extract it. Often, however, role fillers occur in clauses that are not directly linked to an event word. We present a new model for event extraction that jointly considers both the local context around a phrase along with the wider sentential context in a probabilistic framework. Our approach uses a sentential event recognizer and a plausible role-filler recognizer that is conditioned on event sentences. We evaluate our system on two IE data sets and show that our model performs well in comparison to existing IE systems that rely on local phrasal context.

Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm

Jingjing Liu and Stephanie Seneff

This paper presents a parse-and-paraphrase paradigm to assess the degrees of sentiment for product reviews. Sentiment identification has been well studied; however, most previous work provides binary polarities only (positive and negative), and the polarity of sentiment is simply reversed when a negation is detected. The extraction of lexical features such as unigram/bigram also complicates the sentiment classification task, as linguistic structure such as implicit long-distance dependency is often disregarded. In this paper, we propose an approach to extracting adverb-adjective-noun phrases based on clause structure obtained by parsing sentences into a hierarchical representation. We also propose a robust general solution for modeling the contribution of adverbials and negation to the score for degree of sentiment. In an application involving extracting aspect-based pros and cons from restaurant reviews, we obtained a 45% relative improvement in recall through the use of parsing methods, while also improving precision.

Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification

Swapna Somasundaran, Galileo Namata, Janyce Wiebe and Lise Getoor

This work investigates design choices in modeling discourse relations for improving opinion polarity classification. We explore ways in which a discourse-based scheme and its annotations can be exploited to augment and improve a local-information based polarity classifier. For this, two diverse global inference paradigms are used: a supervised collective classification framework that facilitates machine learning using relational information and an unsupervised optimization framework which expresses the intuitions behind the discourse relations as constraints. Both our approaches perform substantially better than the classifier trained on local information alone, establishing the efficacy of both, our methods and the underlying discourse scheme. Finally, we present quantitative and qualitative analyses to show how our approach achieves improvements.

Sentiment Analysis of Conditional Sentences

Ramanathan Narayanan, Bing Liu and Alok Choudhary

This paper studies sentiment analysis of conditional sentences. The aim is to determine whether opinions expressed on different topics in a conditional sentence are positive, negative or neutral. Conditional sentences are one of the commonly used language constructs in text. In a typical document, there are around 8% of such sentences. Due to the condition clause, sentiments expressed in a conditional sentence can be hard to determine. For example, in the sentence, if your Nokia phone is not good, buy this great Samsung phone, the author is positive about "Samsung phone" but does not express an opinion on "Nokia phone" (although the owner of the "Nokia phone" may be negative about it). However, if the sentence does not have "if", the first clause is clearly negative. Although "if" commonly signifies a conditional sentence, there are many other words and constructs that can express conditions. This paper first presents a linguistic analysis of such sentences, and then builds some supervised learning models to determine if sentiments expressed on different topics in a conditional sentence are positive, negative or neutral. Experimental results on conditional sentences from 5 diverse domains are given to demonstrate the effectiveness of the proposed approach.

Subjectivity Word Sense Disambiguation

Cem Akkaya, Janyce Wiebe and Rada Mihalcea

This paper investigates a new task, subjectivity word sense disambiguation (SWSD), which is to automatically determine which word instances in a corpus are being used with subjective senses, and which are being used with objective senses. We provide empirical evidence that SWSD is more feasible than full word sense disambiguation, and that it can be exploited to improve the performance of contextual subjectivity and sentiment analysis systems.

Non-projective parsing for statistical machine translation

Xavier Carreras and Michael Collins

We describe a novel approach for syntax-based statistical MT, which builds on a variant of tree adjoining grammar (TAG). Inspired by work in discriminative dependency parsing, the key idea in our approach is to allow highly flexible reordering operations during parsing, in combination with a discriminative model that can condition on rich features of the source-language string. Experiments on translation from German to English show improvements over phrase-based systems, both in terms of BLEU scores and in human evaluations.

Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models

Arne Mauser, Saša Hasan and Hermann Ney

In this work, we propose two extensions of standard word lexicons in statistical machine translation: A discriminative word lexicon that uses sentence-level source information to predict the target words and a trigger-based lexicon model that extends IBM model 1 with a second trigger, allowing for a more fine-grained lexical choice of target words. The models capture dependencies that go beyond the scope of conventional SMT models such as phrase- and language models. We show that the models improve translation quality by 1% in BLEU over a competitive baseline on a large-scale translation task.

Feature-Rich Translation by Quasi-Synchronous Lattice Parsing

Kevin Gimpel and Noah A. Smith

We present a machine translation framework that can incorporate arbitrary features of both input and output sentences. The core of the approach is a novel decoder based on lattice parsing with quasi-synchronous grammar (Smith and Eisner, 2006), a syntactic formalism that does not require source and target trees to be isomorphic. Using generic approximate dynamic programming techniques, this decoder can handle "non-local" features. Similar approximate inference techniques support efficient parameter estimation with hidden variables. We use the decoder to conduct controlled experiments on a German-to-English translation task, to compare lexical phrase, syntax, and combined models, and to measure effects of various restrictions on non-isomorphism.

Improved Word Alignment with Statistics and Linguistic Heuristics

Ulf Hermjakob

We present a method to align words in a bitext that combines elements of a traditional statistical approach with linguistic knowledge. We demonstrate this approach for Arabic-English, using an alignment lexicon produced by a statistical word aligner, as well as linguistic resources ranging from an English parser to heuristic alignment rules for function words. These linguistic heuristics have been generalized from a development corpus of 100 parallel sentences. Our aligner, UALIGN, outperforms both the commonly used GIZA++ aligner and the state-of-the-art LEAF aligner on F-measure and produces superior scores in end-to-end statistical machine translation, +1.3 BLEU points over GIZA++, and +0.7 over LEAF.

Entity Extraction via Ensemble Semantics

Marco Pennacchiotti and Patrick Pantel

Combining information extraction systems yields significantly higher quality resources than each system in isolation. In this paper, we generalize such a mixing of sources and features in a framework called Ensemble Semantics. We show very large gains in entity extraction by combining state-of-the-art distributional and pattern-based systems with a large set of features from a webcrawl, query logs, and Wikipedia. Experimental results on a webscale extraction of actors, athletes and musicians show significantly higher mean average precision scores (29% gain) compared with the current state of the art.

Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora

Daniel Ramage, David Hall, Ramesh Nallapati and Christopher D. Manning

A significant portion of the world's text is tagged by readers on social bookmarking websites. Credit attribution is an inherent problem in these corpora because most pages have multiple tags, but the tags do not always apply with equal specificity across the whole document. Solving the credit attribution problem requires associating each word in a document with the most appropriate tags and vice versa. This paper introduces Labeled LDA, a topic model that constrains Latent Dirichlet Allocation by defining a one-to-one correspondence between LDA's latent topics and user tags. This allows Labeled LDA to directly learn word-tag correspondences. We demonstrate Labeled LDA's improved expressiveness over traditional LDA with visualizations of a corpus of tagged web pages from *del.icio.us*. Labeled LDA outperforms SVMs by more than 3 to 1 when extracting tag-specific document snippets. As a multi-label text classifier, our model is competitive with a discriminative baseline on a variety of datasets.

Clustering to Find Exemplar Terms for Keyphrase Extraction

Zhiyuan Liu, Peng Li, Yabin Zheng and Maosong Sun

Keyphrases are widely used as a brief summary of documents. Since manual assignment is time-consuming, various unsupervised ranking methods based on importance scores are proposed for keyphrase extraction. In practice, the keyphrases of a document should not only be statistically important in the document, but also have a good coverage of the document. Based on this observation, we propose an unsupervised method for keyphrase extraction. Firstly, the method finds exemplar terms by leveraging clustering techniques, which guarantees the document to be semantically covered by these exemplar terms. Then the keyphrases are extracted from the document using the exemplar terms. Our method outperforms sate-of-the-art graph-based ranking methods (TextRank) by 9.5% in F1-measure.

Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns

Dmitry Davidov and Ari Rappoport

One of the most desired information types when planning a trip to some place is the knowledge of transport, roads and geographical connectedness of prominent sites in this place. While some transport companies or repositories make some of this information accessible, it is not easy to find, and the majority of information about uncommon places can only be found in web free text such as blogs and forums. In this paper we present an algorithmic framework which allows an automated acquisition of map-like information from the web, based on surface patterns like "from X to Y". Given a set of locations as initial seeds, we retrieve from the web an extended set of locations and produce a map-like network which connects these locations using transport type edges. We evaluate our framework in several settings, producing meaningful and precise connection sets.

Wikipedia as frame information repository

Sara Tonelli and Claudio Giuliano

In this paper, we address the issue of automatic extending lexical resources by exploiting existing knowledge repositories. In particular, we deal with the new task of linking FrameNet and Wikipedia using a word sense disambiguation system that, for a given pair framelexical unit (F, l), finds the Wikipage that best expresses the the meaning of l. The mapping can be exploited to straightforwardly acquire new example sentences and new lexical units, both for English and for all languages available in Wikipedia. In this way, it is possible to easily acquire good-quality data as a starting point for the creation of FrameNet in new languages. The evaluation reported both for the monolingual and the multilingual expansion of FrameNet shows that the approach is promising.

Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk

Chris Callison-Burch

Manual evaluation of translation quality is generally thought to be excessively time consuming and expensive. We explore a fast and inexpensive way of doing it using Amazon's Mechanical Turk to pay small sums to a large number of non-expert annotators. For $10 we redundantly recreate judgments from a WMT08 translation task. We find that when combined non-expert judgments have a high-level of agreement with the existing gold-standard judgments of machine translation quality, and correlate more strongly with expert judgments than Bleu does. We go on to show that Mechanical Turk can be used to calculate human-mediated translation edit rate (HTER), to conduct reading comprehension experiments with machine translation, and to create high quality reference translations.

How well does active learning actually work? Time-based evaluation of cost-reduction strategies for language documentation.

Jason Baldridge and Alexis Palmer

Machine involvement has the potential to speed up language documentation. We assess this potential with timed annotation experiments that consider annotator expertise, example selection methods, and suggestions from a machine classifier. We find that better example selection and label suggestions improve efficiency, but effectiveness depends strongly on annotator expertise. Our expert performed best with uncertainty selection, but gained little from suggestions. Our non-expert performed best with random selection and suggestions. The results underscore the importance both of measuring annotation cost reductions with respect to time and of the need for cost-sensitive learning methods that adapt to annotators.

Automatically Evaluating Content Selection in Summarization without Human Models

Annie Louis and Ani Nenkova

We present a fully automatic method for content selection evaluation in summarization that does not require the creation of human model summaries. Our work capitalizes on the assumption that the distribution of words in the input and an informative summary of that input should be similar to each other. Results on a large scale evaluation from the Text Analysis Conference show that input-summary comparisons are very effective for the evaluation of content selection. Our automatic methods rank participating systems similarly to manual model-based pyramid evaluation and to manual human judgments of responsiveness. The best feature, Jensen-Shannon divergence, leads to a correlation as high as 0.88 with manual pyramid evaluations and 0.73 with responsiveness judgements.

Classifier Combination for Contextual Idiom Detection Without Labelled Data

Linlin Li and Caroline Sporleder

We propose a novel unsupervised approach for distinguishing literal and non-literal use of idiomatic expressions. Our model combines an unsupervised and a supervised classifier. The former bases its decision on the cohesive structure of the context and labels training data for the latter, which can then take a larger feature space into account. We show that a combination of both classifiers leads to significant improvements over using the unsupervised classifier alone.

Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing

Brian Roark, Asaf Bachrach, Carlos Cardenas and Christophe Pallier

A number of recent publications have made use of the incremental output of stochastic parsers to derive measures of high utility for psycholinguistic modeling, following the work of Hale (2001; 2003; 2006). In this paper, we present novel methods for calculating separate lexical and syntactic surprisal measures from a single incremental parser using a lexicalized PCFG. We also present an approximation to entropy measures that would otherwise be intractable to calculate for a grammar of that size. Empirical results demonstrate the utility of our methods in predicting human reading times.

It's Not You, it's Me: Detecting Flirting and its Misperception in Speed-Dates

Rajesh Ranganath, Dan Jurafsky and Dan McFarland

Automatically detecting human social intentions from spoken conversation is an important task for dialogue understanding. Since the social intentions of the speaker may differ from what is perceived by the hearer, systems that analyze human conversations need to be able to extract both the perceived and the intended social meaning. We investigate this difference between intention and perception by using a spoken corpus of speed-dates in which both the speaker and the listener rated the speaker on flirtatiousness. Our flirtation-detection system uses prosodic, dialogue, and lexical features to detect a speaker's intent to flirt with up to 71.5% accuracy, significantly outperforming the baseline, but also outperforming the human interlocuters. Our system addresses lexical feature sparsity given the small amount of training data by using an autoencoder network to map sparse lexical feature vectors into 30 compressed features. Our analysis shows that humans are very poor perceivers of intended flirtatiousness, instead often projecting their own intended behavior onto their interlocutors.

Recognizing Implicit Discourse Relations in the Penn Discourse Treebank

Ziheng Lin, Min-Yen Kan and Hwee Tou Ng

We present an implicit discourse relation classifier in the Penn Discourse Treebank (PDTB). Our classifier considers the context of the two arguments, word pair information, as well as the arguments' internal constituent and dependency parses. Our results on the PDTB yields a significant 14.1% improvement over the baseline. In our error analysis, we discuss four challenges in recognizing implicit relations in the PDTB.

A Bayesian Model of Syntax-Directed Tree to String Grammar Induction

Trevor Cohn and Phil Blunsom

Tree based translation models are a compelling means of integrating linguistic information into machine translation. Syntax can inform lexical selection and reordering choices and thereby improve translation quality. Research to date has focussed primarily on decoding with such models, but not on the difficult problem of inducing the bilingual grammar from data. We propose a generative Bayesian model of tree-to-string translation which induces grammars that are both smaller and produce better translations than the previous heuristic two-stage approach which employs a separate word alignment step.

Better Synchronous Binarization for Machine Translation

Tong Xiao, Mu Li, Dongdong Zhang, Jingbo Zhu and Ming Zhou

Binarization of Synchronous Context Free Grammars (SCFG) is essential for achieving polynomial time complexity of decoding for SCFG parsing based machine translation systems. In this paper, we first investigate the excess edge competition issue caused by a left-heavy binary SCFG derived with the method of Zhang et al. (2006). Then we propose a new binarization method to mitigate the problem by exploring other alternative equivalent binary SCFGs. We present an algorithm that iteratively improves the resulting binary SCFG, and empirically show that our method can improve a string-to-tree statistical machine translations system based on the synchronous binarization method in Zhang et al. (2006) on the NIST machine translation evaluation tasks.

Accuracy-Based Scoring for DOT: Towards Direct Error Minimization for Data-Oriented Translation

Daniel Galron, Sergio Penkale, Andy Way and I. Dan Melamed

In this work we present a novel technique to rescore fragments in the Data-Oriented Translation model based on their contribution to translation accuracy. We describe three new rescoring methods, and present the initial results of a pilot experiment on a small subset of the Europarl corpus. This work is a proof-of-concept, and is the first step in directly optimizing translation decisions solely on the hypothesized accuracy of potential translations resulting from those decisions.

Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases

Yuval Marton, Chris Callison-Burch and Philip Resnik

Untranslated words still constitute a major problem for Statistical Machine Translation (SMT), and current SMT systems are limited by the quantity of parallel training texts. Augmenting the training data with paraphrases generated by pivoting through other languages alleviates this problem, especially for the so-called "low density" languages. But pivoting requires additional parallel texts. We address this problem by deriving paraphrases monolingually, using distributional semantic similarity measures, thus providing access to larger training resources, such as comparable and unrelated monolingual corpora. We present what is to our knowledge the first successful integration of a collocational approach to untranslated words with an end-to-end, state of the art SMT system demonstrating significant translation improvements in a low-resource setting.

A Comparison of Model Free versus Model Intensive Approaches to Sentence Compression

Tadashi Nomoto

This work introduces a model free approach to sentence compression, which grew out of ideas from Nomoto (2008), and examines how it compares to a state-of-art model intensive approach known as Tree-to-Tree Transducer, or T3 (Cohn and Lapata, 2008). It is found that a model free approach significantly outperforms T3 on the particular data we created from the Internet. We also discuss what might have caused T3's poor performance.

Natural Language Generation with Tree Conditional Random Fields

Wei Lu, Hwee Tou Ng and Wee Sun Lee

This paper presents an effective method for generating natural language sentences from their underlying meaning representations. The method is built on top of a hybrid tree representation that jointly encodes both the meaning representation as well as the natural language in a tree structure. By using a tree conditional random field on top of the hybrid tree representation, we are able to explicitly model phrase-level dependencies amongst neighboring natural language phrases and meaning representation components in a simple and natural way. We show that the additional dependencies captured by the tree conditional random field allows it to perform better than directly inverting a previously developed hybrid tree semantic parser. Furthermore, we demonstrate that the model performs better than a previous state-of-the-art natural language generation model. Experiments are performed on two benchmark corpora with standard automatic evaluation metrics.

Perceptron Reranking for CCG Realization

Michael White and Rajakrishnan Rajkumar

This paper shows that discriminative reranking with an averaged perceptron model yields substantial improvements in realization quality with CCG. The paper confirms the utility of including language model log probabilities as features in the model, which prior work on discriminative training with log linear models for HPSG realization had called into question. The perceptron model allows the combination of multiple n-gram models to be optimized and then augmented with both syntactic features and discriminative n-gram features. The full model yields a state-of-the-art BLEU score of 0.8506 on Section 23 of the CCGbank, to our knowledge the best score reported to date using a reversible, corpus-engineered grammar.

Multi-Document Summarisation Using Generic Relation Extraction

Ben Hachey

Experiments are reported that investigate the effect of various source document representations on the accuracy of the sentence extraction phase of a multi-document summarisation task. A novel representation is introduced based on generic relation extraction (GRE), which aims to build systems for relation identification and characterisation that can be transferred across domains and tasks without modification of model parameters. Results demonstrate performance that is significantly higher than a non-trivial baseline that uses tf*idf-weighted words and is at least as good as a comparable but less general approach from the literature. Analysis shows that the representations compared are complementary, suggesting that extraction performance could be further improved through system combination.

Language Models Based on Semantic Composition

Jeff Mitchell and Mirella Lapata

In this paper we propose a novel statistical language model to capture long-range semantic dependencies. Specifically, we apply the concept of semantic composition to the problem of constructing predictive history representations for upcoming words. We also examine the influence of the underlying semantic space on the composition task by comparing spatial semantic representations against topic-based ones. The composition models yield reductions in perplexity when combined with a standard n-gram language model over the n-gram model alone. We also obtain perplexity reductions when integrating our models with a structured language model.

Graded Word Sense Assignment

Katrin Erk and Diana McCarthy

Word sense disambiguation is typically phrased as the task of labeling a word in context with the best-fitting sense from a sense inventory such as WordNet. While questions have often been raised over the choice of sense inventory, computational linguists have readily accepted the best-fitting sense methodology despite the fact that the case for discrete sense boundaries is widely disputed by lexical semantics researchers. This paper studies graded word sense assignment, based on a recent dataset of graded word sense annotation.

Joint Learning of Preposition Senses and Semantic Roles of Prepositional Phrases

Daniel Dahlmeier, Hwee Tou Ng and Tanja Schultz

The sense of a preposition is related to the semantics of its dominating prepositional phrase. Knowing the sense of a preposition could help to correctly classify the semantic role of the dominating prepositional phrase and vice versa. In this paper, we propose a joint probabilistic model for word sense disambiguation of prepositions and semantic role labeling of prepositional phrases. Our experiments on the PropBank corpus show that jointly learning the word sense and the semantic role leads to an improvement over state-of-the-art individual classifier models on the two tasks.

Projecting Parameters for Multilingual Word Sense Disambiguation

Mitesh M. Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya

We report in this paper a way of doing Word Sense Disambiguation (WSD) that has its origin in multilingual MT and that is cognizant of the fact that parallel corpora, wordnets and sense annotated corpora are scarce resources. With respect to these resources, languages show different levels of readiness; however a more resource fortunate language can help a less resource fortunate language. Our WSD method can be applied to a language even when no sense tagged corpora for that language is available. This is achieved by projecting wordnet and corpus parameters from another language to the language in question. The approach is centered around a novel synset based multilingual dictionary and the empirical observation that within a domain the distribution of senses remains more or less invariant across languages. The effectiveness of our approach is verified by doing parameter projection and then running two different WSD algorithms. The accuracy values of approximately 75% (F1-score) for three languages in two different domains establish the fact that within a domain it is possible to circumvent the problem of scarcity of resources by projecting parameters like sense distributions, corpus co-occurrences, conceptual distance, etc. from one language to another.

Poster Presentations – Thursday, 6 August

Gazpacho and summer rash: lexical relationships from temporal patterns of web search queries

Enrique Alfonseca, Massimiliano Ciaramita and Keith Hall

In this paper we investigate temporal patterns of web search queries. We carry out several evaluations to analyze the properties of temporal profiles of queries, revealing promising semantic and pragmatic relationships between words. We focus on two applications: query suggestion and query categorization. The former shows a potential for time-series similarity measures to identify specific semantic relatedness between words, which results in state-of-the-art performance in query suggestion while providing complementary information to more traditional distributional similarity measures. The query categorization evaluation suggests that the temporal profile alone is not a strong indicator of broad topical categories.

A Compact Forest for Scalable Inference over Entailment and Paraphrase Rules

Roy Bar-Haim, Jonathan Berant and Ido Dagan

A large body of recent research has been investigating the acquisition and application of applied inference knowledge. Such knowledge may be typically captured as entailment rules, applied over syntactic representations. Efficient inference with such knowledge then becomes a fundamental problem. Starting out from a formalism for entailment-rule application we present a novel packed data-structure and a corresponding algorithm for its scalable implementation. We proved the validity of the new algorithm and established its efficiency analytically and empirically.

Discriminative Substring Decoding for Transliteration

Colin Cherry and Hisami Suzuki

We present a discriminative substring decoder for transliteration. This decoder extends recent approaches for discriminative character transduction by allowing for a list of known target-language words, an important resource for transliteration. Our approach improves upon Sherif and Kondrak's (2007b) state-of-the-art decoder, creating a 28.5% relative improvement in transliteration accuracy on a Japanese katakana-to-English task. We also conduct a controlled comparison of two feature paradigms for discriminative training: indicators and hybrid generative features. Surprisingly, the generative hybrid outperforms its purely discriminative counterpart, despite losing access to rich source-context features. Finally, we show that machine transliterations have a positive impact on machine translation quality, improving human judgments by 0.5 on a 4-point scale.

Re-Ranking Models Based-on Small Training Data for Spoken Language Understanding

Marco Dinarelli, Alessandro Moschitti and Giuseppe Riccardi

The design of practical language applications by means of statistical approaches requires annotated data, which is one of the most critical constraint. This is particularly true for Spoken Dialog Systems since considerably domain-specific conceptual annotation is needed to obtain accurate Language Understanding models. Since data annotation is usually costly, methods to reduce the amount of data are needed. In this paper, we show that better feature representations serve the above purpose and that structure kernels provide the needed improved representation. Given the relatively high computational cost of kernel methods, we apply them to just re-rank the list of hypotheses provided by a fast generative model. Experiments with Support Vector Machines and different kernels on two different dialog corpora show that our re-ranking models can achieve better results than state-of-the-art approaches when small data is available.

Empirical Exploitation of Click Data for Task Specific Ranking

Anlei Dong, Yi Chang, Shihao Ji, Ciya Liao, Xin Li and Zhaohui Zheng

There have been increasing needs for task specific rankings in web search such as rankings for specific query segments like long queries, time-sensitive queries, navigational queries, etc; or rankings for specific domains/contents like answers, blogs, news, etc. In the spirit of divide-and-conquer, task specific ranking may have potential advantages over generic ranking since different tasks have task-specific features, data distributions, as well as feature grade correlations. A critical problem for the task-specific ranking is training data insufficiency, which may be solved by using the data extracted from click log. This paper empirically studies how to appropriately exploit click data to improve rank function learning in task-specific ranking. The main contributions are 1) the exploration on the utilities of two promising approaches for click pair extraction; 2) the analysis of the role played by the noise information which inevitably appears in click data extraction; 3) the appropriate strategy for combining training data and click data; 4) the comparison of click data which are consistent and inconsistent with baseline function.

The Feature Subspace Method for SMT System Combination

Nan Duan, Mu Li, Tong Xiao and Ming Zhou

Recently system combination has been shown to be an effective way to improve translation quality over single machine translation systems. In this paper, we present a simple and effective method to systematically derive an ensemble of SMT systems from one baseline linear SMT model for use in system combination. Each system in the resulting ensemble is based on a feature set derived from the features of the baseline model (typically a subset of it). We will discuss the principles to determine the feature sets for derived systems, and present in detail the system combination model used in our work. Evaluation is performed on the data sets for NIST 2004 and NIST 2005 Chinese-to-English machine translation tasks. Experimental results show that our method can bring significant improvements to baseline systems with state-of-the-art performance.

Lattice-based System Combination for Statistical Machine Translation

Yang Feng, Yang Liu, Haitao Mi, Qun Liu and Yajuan Lü

Current system combination methods usually use confusion networks to find consensus translations among different systems. Requiring one-to-one mappings between the words in candidate translations, confusion networks have difficulty in handling more general situations in which several words are connected to another several words. Instead, we propose a lattice-based system combination model that allows for such phrase alignments and uses *lattice*s to encode all candidate translations. Experiments show that our approach achieves significant improvements over the state-of-the-art baseline system on Chinese-to-English translation test sets.

A Joint Language Model With Fine-grain Syntactic Tags

Denis Filimonov and Mary Harper

We present a scalable joint language model designed to utilize fine-grain syntactic tags. We discuss challenges such a design faces and describe our solutions that scale well to large tagsets and corpora. We advocate the use of relatively simple tags that do not require deep linguistic knowledge of the language but provide more structural information than POS tags and can be derived from automatically generated parse trees – a combination of properties that allows easy adoption of this model for new languages. We propose two fine-grain tagsets and evaluate our model using these tags, as well as POS tags and SuperARV tags in a speech recognition task and discuss future directions.

Bidirectional Phrase-based Statistical Machine Translation

Andrew Finch and Eiichiro Sumita

This paper investigates the effect of direction in phrase-based statistial machine translation decoding. We compare a typical phrase-based machine translation decoder using a left-to-right decoding strategy to a right-to-left decoder. We also investigate the effectiveness of a bidirectional decoding strategy that integrates both mono-directional approaches, with the aim of reducing the effects due to language specificity. Our experimental evaluation was extensive, based on 272 different language pairs, and gave the surprising result that for most of the language pairs, it was better decode from right-to-left than from left-to-right. As expected the relative performance of left-to-right and right-to-left strategies proved to be highly language dependent. The bidirectional approach outperformed the both the left-to-right strategy and the right-to-left strategy, showing consistent improvements that appeared to be unrelated to the specific languages used for translation. Bidirectional decoding gave rise to an improvement in performance over a left-to-right decoding strategy in terms of the BLEU score in 99% of our experiments.

Real-time decision detection in multi-party dialogue

Matthew Frampton, Jia Huang, Trung Bui and Stanley Peters

We describe a process for automatically detecting decision-making sub-dialogues in multi-party, human-human meetings in real-time. Our basic approach to decision detection involves distinguishing between different utterance types based on the roles that they play in the formulation of a decision. We use what we refer to as “hierarchical classification”. Here, independent binary sub-classifiers detect the different decision dialogue acts, and then based on the sub-classifier hypotheses, a super-classifier determines which dialogue regions are decision discussions. In this paper, we describe how this approach can be implemented in real-time, and show that the resulting system’s performance compares well with that of other decision detectors, including an off-line detector. In testing on a portion of the AMI corpus, the real-time systemachieves an F-score of .54, and the off-line detector, .55, and this difference is not statistically significant.

On the Role of Lexical Features in Sequence Labeling

Yoav Goldberg and Michael Elhadad

We use the technique of SVM anchoring to demonstrate that lexical features extracted from a training corpus are not necessary to obtain state of the art results on tasks such as Named Entity Recognition and Chunking. While standard models require as many as 100K distinct features, we derive models with as little as 1K features that perform as well or better on different domains. These robust reduced models indicate that the way rare lexical features contribute to classification in NLP is not fully understood. Contrastive error analysis (with and without lexical features) indicates that lexical features do contribute to resolving some semantic and complex syntactic ambiguities – but we find this contribution does not generalize outside the training corpus. As a general strategy, we believe lexical features should not be directly derived from a training corpus but instead, carefully inferred and selected from other sources.

Simple Coreference Resolution with Rich Syntactic and Semantic Features

Aria Haghighi and Dan Klein

Coreference systems are driven by syntactic, semantic, and discourse constraints. We present a simple approach which completely modularizes these three aspects. In contrast to much current work, which focuses on learning and on the discourse component, our system is deterministic and is driven entirely by syntactic and semantic compatibility as learned from a large, unlabeled corpus. Despite its simplicity and discourse naivete, our system substantially outperforms all unsupervised systems and most supervised ones. Primary contributions include (1) the presentation of a simple-to-reproduce, high-performing baseline and (2) the demonstration that most remaining errors can be attributed to syntactic and semantic factors external to the coreference phenomenon (and perhaps best addressed by non-coreference systems).

Descriptive and Empirical Approaches to Capturing Underlying Dependencies among Parsing Errors

Tadayoshi Hara, Yusuke Miyao and Jun’ichi Tsujii

In this paper, we provide descriptive and empirical approaches to effectively extracting underlying dependencies among parsing errors. In the descriptive approach, we define some combinations of error patterns and extract them from given errors. In the empirical approach, on the other hand, we re-parse a sentence with a target error corrected and observe errors corrected together. Experiments on an HPSG parser show that each of these approaches can clarify the dependencies among individual errors from each point of view. Moreover, the comparison between the results of the two approaches shows that combining these approaches can achieve a more detailed error analysis.

Large-Scale Verb Entailment Acquisition from the Web

Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Stijn De Saeger, Masaki Murata and Jun’ichi Kazama

Textual entailment recognition plays a fundamental role in tasks that require in-depth natural language understanding. In order to use entailment recognition technologies for real-world applications, a large-scale entailment knowledge base is indispensable. This paper proposes a conditional probability based directional similarity measure to acquire verb entailment pairs on a large scale. We targeted 52,562 verb types that were derived from 100,000,000 Japanese Web documents, without regard for whether they were used in daily life or only in specific fields. In an evaluation of the top 20,000 verb entailment pairs acquired by previous methods and ours, we found that our similarity measure outperformed the previous ones. Our method also worked well for the top 100,000 results.

A Syntactified Direct Translation Model with Linear-time Decoding

Hany Hassan, Khalil Sima'an and Andy Way

Recent syntactic extensions of statistical translation models work with a (Context-Free or Tree-Substitution) synchronous grammar extracted from an automatically parsed parallel corpus. The decoders accompanying these extensions typically exceed quadratic time complexity. This paper extends the Direct Translation Model 2 (DTM2) with syntax while maintaining linear-time decoding. We employ a linear-time parsing algorithm based on an eager, incremental interpretation of Combinatory Categorial Grammar (CCG). As every input word is processed, the local parsing decisions resolve ambiguity eagerly, by selecting a single supertag–operator pair for extending the dependency parse incrementally. Alongside translation features extracted from the derived parse tree, we explore syntactic features extracted from the incremental derivation process. Our empirical experiments show that our model significantly outperforms the state-of-the art DTM2 system and a standard phrase-based translation system.

Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge

Samer Hassan and Rada Mihalcea

In this paper, we address the task of cross-lingual semantic relatedness. We introduce a method that relies on the information extracted from Wikipedia, by exploiting the inter-language links available between Wikipedia versions in multiple languages. Through experiments performed on several language pairs, we show that the method performs well, with a performance comparable to monolingual measures of relatedness.

Joint Optimization for Machine Translation System Combination

Xiaodong He and Kristina Toutanova

System combination has emerged as a powerful method for machine translation (MT). This paper pursues a joint optimization strategy for combining outputs from multiple MT systems, where word alignment, ordering, and lexical selection decisions are made jointly according to a set of feature functions combined in a single log-linear model. The decoding algorithm is described in detail and a set of new features that support this joint decoding approach is proposed. The approach is evaluated in comparison to state-of-the-art confusion-network-based system combination methods using equivalent features and shown to outperform them significantly.

Fully Lexicalising CCGbank with Hat Categories

Matthew Honnibal and James R. Curran

We introduce an extension to CCG that allows form and function to be represented simultaneously, reducing the proliferation of modifier categories seen in standard CCG analyses. We can then remove the non-combinatory rules CCGbank uses to address this problem, producing a grammar that is fully lexicalised and far less ambiguous. There are intrinsic benefits to full lexicalisation, such as semantic transparency and simpler domain adaptation. The clearest advantage is a 52-88% improvement in parse speeds, which comes with only a small reduction in accuracy.

Bilingually-Constrained (Monolingual) Shift-Reduce Parsing

Liang Huang, Wenbin Jiang and Qun Liu

Jointly parsing two languages has been shown to improve accuracies on either or both sides. However, its search space is much bigger than the monolingual case, forcing existing approaches to employ complicated modeling and crude approximations. Here we propose a much simpler alternative, *bilingually-constrained monolingual parsing*, where a source-language parser learns to exploit reorderings between languages as additional observation, but *not* bothering to build the target-side tree as well. We show specifically how to enhance a shift-reduce dependency parser with alignment features to resolve shift-reduce conflicts. Experiments on the bilingual portion of Chinese Treebank show that, with just 3 bilingual features, we can improve parsing accuracies by 0.6% (absolute) for both English and Chinese over a state-of-the-art baseline, with negligible ($\sim$6%) efficiency overhead, thus much faster than biparsing.

Accurate Semantic Class Classifier for Coreference Resolution

Zhiheng Huang, Guangping Zeng, Weiqun Xu and Asli Celikyilmaz

There have been considerable attempts to incorporate semantic knowledge into coreference resolution systems: different knowledge sources such as WordNet and Wikipedia have been used to boost the performance. In this paper, we propose new ways to extract WordNet feature. This feature, along with other features such as named entity feature, can be used to build an accurate semantic class (SC) classifier. In addition, we analyze the SC classification errors and propose to use relaxed SC agreement features. The proposed accurate SC classifier and the relaxation of SC agreement features on ACE2 coreference evaluation can boost our baseline system by 10.4% and 9.7% using MUC score and anaphor accuracy respectively.

Real-Word Spelling Correction using Google Web 1T 3-grams

Aminul Islam and Diana Inkpen

We present a method for detecting and correcting multiple real-word spelling errors using the Google Web 1T 3-gram data set and a normalized and modified version of the Longest Common Subsequence (LCS) string matching algorithm. Our method is focused mainly on how to improve the detection recall (the fraction of errors correctly detected) and the correction recall (the fraction of errors correctly amended), while keeping the respective precisions (the fraction of detections or amendments that are correct) as high as possible. Evaluation results on a standard data set show that our method outperforms two other methods on the same task.

Semi-supervised Speech Act Recognition in Emails and Forums

Minwoo Jeong, Chin-Yew Lin and Gary Geunbae Lee

In this paper, we present a semi-supervised method for automatic speech act recognition in email and forums. The major challenge of this task is due to lack of labeled data in these two genres. Our method leverages labeled data in the Switchboard-DAMSL and the Meeting Recorder Dialog Act database and applies simple domain adaptation techniques over a large amount of unlabeled email and forum data to address this problem. Our method uses automatically extracted features such as phrases and dependency trees, called subtree features, for semi-supervised learning. Empirical results demonstrate that our model is effective in email and forum speech act recognition.

Using Morphological and Syntactic Structures for Chinese Opinion Analysis

Lun-Wei Ku, Ting-Hao Huang and Hsin-Hsi Chen

This paper employs morphological struc-tures and relations between sentence seg-ments for opinion analysis on words and sentences. Chinese words are classified into eight morphological types by two proposed classifiers, CRF classifier and SVM classifier. Experiments show that the injection of morphological information improves the performance of the word po-larity detection. To utilize syntactic struc-tures, we annotate structural trios to repre-sent relations between sentence segments. Experiments show that considering struc-tural trios is useful for sentence opinion analysis. The best f-score achieves 0.77 for opinion word extraction, 0.62 for opin-ion word polarity detection, 0.80 for opin-ion sentence extraction, and 0.54 for opin-ion sentence polarity detection.

Finding Short Definitions of Terms on Web Pages

Gerasimos Lampouras and Ion Androutsopoulos

We present a system that finds short definitions of terms on Web pages. It employs a Maximum Entropy classifier, but it is trained on automatically generated examples; hence, it is in effect unsupervised. We use ROUGE-W to generate training examples from encyclopedias and Web snippets, a method that outperforms an alternative centroid-based one. After training, our system can be used to find definitions of terms that are not covered by encyclopedias. The system outperforms a comparable publicly available system, as well as a previously published form of our system.

Improving Nominal SRL in Chinese Language with Verbal SRL Information and Automatic Predicate Recognition

Junhui Li, GuoDong Zhou, Hai Zhao, Qiaoming Zhu and Peide Qian

This paper explores Chinese semantic role la-beling (SRL) for nominal predicates. Besides those widely used features in verbal SRL, various nominal SRL-specific features are first included. Then, we improve the performance of nominal SRL by integrating useful features derived from a state-of-the-art verbal SRL sys-tem. Finally, we address the issue of automatic predicate recognition, which is essential for a nominal SRL system. Evaluation on Chinese NomBank shows that our pioneer research in integrating various features derived from ver-bal SRL significantly improves the perform-ance. It also shows that our nominal SRL system much outperforms the state-of-the-art ones.

On the Use of Virtual Evidence in Conditional Random Fields

Xiao Li

Virtual evidence (VE), first introduced by (Pearl, 1988), provides a convenient way of incorporating prior knowledge into Bayesian networks. This work generalizes the use of VE to undirected graphical models and, in particular, to conditional random fields (CRFs). We show that VE can be naturally encoded into a CRF model as potential functions. More importantly, we propose a novel semi-supervised machine learning objective for estimating a CRF model integrated with VE. The objective can be optimized using the Expectation-Maximization algorithm while maintaining the discriminative nature of CRFs. When evaluated on the CLASSIFIEDS data, our approach significantly outperforms the best known solutions reported on this task.

Refining Grammars for Parsing with Hierarchical Semantic Knowledge

Xiaojun Lin, Yang Fan, Meng Zhang, Xihong Wu and Huisheng Chi

This paper proposes a novel method to refine the grammars in parsing by utilizing semantic knowledge from HowNet. Based on the hierarchical state-split approach, which can refine grammars automatically in a data-driven manner, this study introduces semantic knowledge into the splitting process at two steps. Firstly, each part-of-speech node will be annotated with a semantic tag of its terminal word. These new tags generated in this step are semantic-related, which can provide a good start for splitting. Secondly, a knowledge-based criterion is used to supervise the hierarchical splitting of these semantic-related tags, which can alleviate overfitting. The experiments are carried out on both Chinese and English Penn Treebank show that the refined grammars with semantic knowledge can improve parsing performance significantly. Especially with respect to Chinese, our parser achieves an F1 score of 87.5% which is the best published result we are aware of. The further analysis on the refined grammars shows that, our method tends to split the content categories more often than the baseline method and the function classes less often.

Bayesian Learning of Phrasal Tree-to-String Templates

Ding Liu and Daniel Gildea

We examine the problem of overcoming noisy word-level alignments when learning tree-to-string translation rules. Our approach introduces new rules, and reestimates rule probabilities using EM. The major obstacles to this approach are the very reasons that word-alignments are used for rule extraction: the huge space of possible rules, as well as controlling overfitting. By carefully controlling which portions of the original alignments are reanalyzed, and by using Bayesian inference during re-analysis, we show significant improvement over the baseline rules extracted from word-level alignments.

Human-competitive tagging using automatic keyphrase extraction

Olena Medelyan, Eibe Frank and Ian H. Witten

This paper connects two research areas: automatic tagging on the web and statistical keyphrase extraction. First, we analyze the quality of tags in a collaboratively created folksonomy using traditional evaluation techniques. Next, we demonstrate how documents can be tagged automatically with a state-of-the-art keyphrase extraction algorithm, and further improve performance in this new domain using a new algorithm, Maui, that utilizes semantic information extracted from Wikipedia. Maui outperforms existing approaches and extracts tags that are competitive with those assigned by the best performing human taggers.

Supervised Learning of a Probabilistic Lexicon of Verb Semantic Classes

Yusuke Miyao and Jun'ichi Tsujii

The work presented in this paper explores a supervised method for learning a probabilistic model of a lexicon of VerbNet classes. We intend for the probabilistic model to provide a probability distribution of verb-class associations, over known and unknown verbs, including polysemous words. In our approach, training instances are obtained from an existing lexicon and/or from an annotated corpus, while the features, which represent syntactic frames, semantic similarity, and selectional preferences, are extracted from unannotated corpora. Our model is evaluated in type-level verb classification tasks: we measure the prediction accuracy of VerbNet classes for unknown verbs, and also measure the dissimilarity between the learned and observed probability distributions. We empirically compare several settings for model learning, while we vary the use of features, source corpora for feature extraction, and disambiguated corpora. In the task of verb classification into all VerbNet classes, our best model achieved a 10.69% error reduction in the classification accuracy, over the previously proposed model.

A Study on the Semantic Relatedness of Query and Document Terms in Information Retrieval

Christof Müller and Iryna Gurevych

The use of lexical semantic knowledge in information retrieval has been a field of active study for a long time. Collaborative knowledge bases like Wikipedia and Wiktionary, which have been applied in computational methods only recently, offer new possibilities to enhance information retrieval. In order to find the most beneficial way to employ these resources, we analyze the lexical semantic relations that hold among query and document terms and compare how these relations are represented by a measure for semantic relatedness. We explore the potential of different indicators of document relevance that are based on semantic relatedness and compare the characteristics and performance of the knowledge bases Wikipedia, Wiktionary and WordNet.

Predicting Subjectivity in Multimodal Conversations

Gabriel Murray and Giuseppe Carenini

In this research we aim to detect subjective sentences in multimodal conversations. We introduce a novel technique wherein subjective patterns are learned from both labeled and unlabeled data, using n-gram word sequences with varying levels of lexical instantiation. Applying this technique to meeting speech and email conversations, we gain significant improvement over state-of-the-art approaches. Furthermore, we show that coupling the pattern-based approach with features that capture characteristics of general conversation structure yields additional improvement.

Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages

Preslav Nakov and Hwee Tou Ng

We propose a novel language-independent approach for improving statistical machine translation for resource-poor languages by exploiting their similarity to resource-rich ones. More precisely, we improve the translation from a resource-poor source language X1 into a resource-rich language Y given a bi-text containing a limited number of parallel sentences for X1-Y and a larger bi-text for X2-Y for some resource-rich language X2 that is closely related to X1. The evaluation for Indonesian-English (using Malay) and Spanish-English (using Portuguese and pretending Spanish is resource-poor) shows an absolute gain of up to 1.35 and 3.37 Bleu points, respectively, which is an improvement over the rivaling approaches, while using much less additional data.

What's in a name? In some languages, grammatical gender

Vivi Nastase and Marius Popescu

This paper presents an investigation of the relation between the way words sound and their gender in two gendered languages: German and Romanian. Gender is an issue that has long preoccupied linguists and baffled language learners. We verify the hypothesis that gender is dictated by the general sound patterns of a language, and that it goes beyond suffixes or word endings. Experimental results on German and Romanian nouns show strong support for this hypothesis, as we are able to learn a model that predicts the gender of a word based on its form with high accuracy.

Convolution Kernels on Constituent, Dependency and Sequential Structures for Relation Extraction

Truc-Vien T. Nguyen, Alessandro Moschitti and Giuseppe Riccardi

This paper explores the use of innovative kernels based on syntactic and semantic structures for a target relation extraction task. Syntax is derived from constituent and dependency parse trees whereas semantics concerns to entity types and lexical sequences. We investigate the effectiveness of such representations in the automated relation extraction from texts. We process the above data by means of Support Vector Machines along with the syntactic tree, the partial tree and the word sequence kernels. Our study on the ACE 2004 corpus illustrates that the combination of the above kernels achieves high effectiveness and significantly improves the current state-of-the-art.

Automatic Acquisition of the Argument-Predicate Relations from a Frame-Annotated Corpus

Ekaterina Ovchinnikova, Theodore Alexandrov and Tonio Wandmacher

This paper presents an approach to automatic acquisition of the argument-predicate relations from a semantically annotated corpus. We use SALSA, a German newspaper corpus manually annotated with role-semantic information based on frame semantics. Since the relatively small size of SALSA does not allow to estimate the semantic relatedness in the extracted argument-predicate pairs, we use a larger corpus for ranking. Two experiments have been performed in order to evaluate the proposed approach. In the first experiment we compare automatically extracted argument-predicate relations with the gold standard formed from associations provided by human subjects. In the second experiment we calculate correlation between automatic relatedness measure and human ranking of the extracted relations.

Detecting Speculations and their Scopes in Scientific Text

Arzucan Özgür and Dragomir R. Radev

Distinguishing speculative statements from factual ones is important for most biomedical text mining applications. We introduce an approach which is based on solving two sub-problems to identify speculative sentence fragments. The first sub-problem is identifying the speculation keywords in the sentences and the second one is resolving their linguistic scopes. We formulate the first sub-problem as a supervised classification task, where we classify the potential keywords as real speculation keywords or not by using a diverse set of linguistic features that represent the contexts of the keywords. After detecting the actual speculation keywords, we use the syntactic structures of the sentences to determine their scopes.

Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models

Michael Paul and Roxana Girju

This paper presents preliminary results on the detection of cultural differences from people's experiences in various countries from two perspectives: tourists and locals. Our approach is to develop probabilistic models that would provide a good framework for such studies. Thus, we propose here a new model, ccLDA, which extends over the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and crosscollection mixture (ccMix) (Zhai et al., 2004) models on blogs and forums. We also provide a qualitative and quantitative analysis of the model on the cross-cultural data.

Consensus Training for Consensus Decoding in Machine Translation

Adam Pauls, John DeNero and Dan Klein

We propose a novel objective function for discriminatively tuning log-linear machine translation models. Our objective explicitly optimizes the BLEU score of *expected* n-gram counts, the same quantities that arise in forest-based consensus and minimum Bayes risk decoding methods. Our continuous objective can be optimized using simple gradient ascent. However, computing critical quantities in the gradient necessitates a novel dynamic program, which we also present here. Assuming BLEU as an evaluation measure, our objective function has two principle advantages over standard max BLEU tuning. First, it specifically optimizes model weights for downstream consensus decoding procedures. An unexpected second benefit is that it reduces overfitting, which can improve test set BLEU scores when using standard Viterbi decoding.

Using Word-Sense Disambiguation Methods to Classify Web Queries by Intent

Emily Pitler and Ken Church

Three methods are proposed to classify queries by intent (CQI), e.g., navigational, informational, commercial, etc. Following mixed-initiative dialog systems, search engines should distinguish navigational queries where the user is taking the initiative from other queries where there are more opportunities for system initiatives (e.g., suggestions, ads). The query intent problem has a number of useful applications for search engines, affecting how many (if any) advertisements to display, which results to return, and how to arrange the results page. Click logs are used as a substitute for annotation. Clicks on ads are evidence for commercial intent; other types of clicks are evidence for other intents. We start with a simple Naive Bayes baseline that works well when there is plenty of training data. When training data is less plentiful, we back off to nearby URLs in a click graph, using a method similar to Word-Sense Disambiguation. Thus, we can infer that "designer trench" is commercial because it is close to "www.saksfifthavenue.com", which is known to be commercial. The baseline method was designed for precision and the backoff method was designed for recall. Both methods are fast and do not require crawling webpages. We recommend a third method, a hybrid of the two, that does no harm when there is plenty of training data, and generalizes better when there isn't, as a strong baseline for the CQI task.

Semi-Supervised Learning for Semantic Relation Classification using Stratified Sampling Strategy

Longhua Qian, Guodong Zhou, Fang Kong and Qiaoming Zhu

This paper presents a new approach to select-ing the initial seed set using stratified sampling strategy in bootstrapping-based semi-supervised learning for semantic relation classification. First, the training data is partitioned into several strata according to relation types/subtypes, then relation instances are randomly sampled from each stratum to form the initial seed set. We also investigate different augmentation strategies in iteratively adding reliable instances to the labeled set, and find that the bootstrapping procedure may stop at a reasonable point to significantly decrease the training time without de-grading too much in performance. Experiments on the ACE RDC 2003 and 2004 corpora show the stratified sampling strategy contributes more than the bootstrapping procedure itself. This suggests that a proper sampling strategy is critical in semi-supervised learning.

Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis

Changqin Quan and Fuji Ren

There is plenty of evidence that emotion analysis has many valuable applications. In this study a blog emotion corpus is constructed for Chinese emotional expression analysis. This corpus contains manual annotation of eight emotional categories (expect, joy, love, surprise, anxiety, sorrow, angry and hate), emotion intensity, emotion holder/target, emotional word/phrase, degree word, negative word, conjunction, rhetoric, punctuation and other linguistic expressions that indicate emotion. Annotation agreement analyses for emotion classes and emotional words and phrases are described. Then, using this corpus, we explore emotion expressions in Chinese and present the analyses on them.

A Probabilistic Model for Associative Anaphora Resolution

Ryohei Sasano and Sadao Kurohashi

This paper proposes a probabilistic model for associative anaphora resolution in Japanese. Associative anaphora is a type of bridging anaphora, in which the anaphor and its antecedent are not coreferent. Our model regards associative anaphora as a kind of zero anaphora and resolves it in the same manner as zero anaphora resolution using automatically acquired lexical knowledge. Experimental results show that our model resolves associative anaphora with good performance and the performance is improved by resolving it simultaneously with zero anaphora.

Quantifier Scope Disambiguation Using Extracted Pragmatic Knowledge: Preliminary Results

Prakash Srinivasan and Alexander Yates

It is well known that pragmatic knowledge is useful and necessary in many difficult language processing tasks, but because this knowledge is difficult to acquire and process automatically, it is rarely used. We present an open information extraction technique for automatically extracting a particular kind of pragmatic knowledge from text, and we show how to integrate the knowledge into a Markov Logic Network model for quantifier scope disambiguation. Our model improves quantifier scope judgments in experiments.

Chinese Semantic Role Labeling with Shallow Parsing

Weiwei Sun, Zhifang Sui, Meng Wang and Xin Wang

Most existing systems for Chinese Semantic Role Labeling (SRL) make use of full syntactic parses. In this paper, we evaluate SRL methods that take partial parses as inputs. We first extend the study on Chinese shallow parsing presented in (Chen et al., 2006) by raising a set of additional features. On the basis of our shallow parser, we implement SRL systems which cast SRL as the classification of syntactic chunks with IOB2 representation for semantic roles (i.e. semantic chunks). Two labeling strategies are presented: 1) directly tagging semantic chunks in one-stage, and 2) identifying argument boundaries as a chunking task and labeling their semantic types as a classification task. For both methods, we present encouraging results, achieving significant improvements over the best reported SRL performance in the literature. Additionally, we put forward a rule-based algorithm to automatically acquire Chinese verb formation, which is empirically shown to enhance SRL.

Discovery of Term Variation in Japanese Web Search Queries

Hisami Suzuki, Xiao Li and Jianfeng Gao

In this paper we address the problem of identifying a broad range of term variations in Japanese web search queries, where these variations pose a particularly thorny problem due to the multiple character types employed in its writing system. Our method extends the techniques proposed for English spelling correction of web queries to handle a wider range of term variants including spelling mistakes, valid alternative spellings using multiple character types, transliterations and abbreviations. The core of our method is a statistical model built on the MART algorithm (Friedman, 2001). We show that both string and semantic similarity features contribute to identifying term variation in web search queries; specifically, the semantic similarity features used in our system are learned by mining user session and click-through logs, and are useful not only as model features but also in generating term variation candidates efficiently. The proposed method achieves 70% precision on the term variation identification task with the recall slightly higher than 60%, reducing the error rate of a nave baseline by 38%.

Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics

Simone Teufel, Advaith Siddharthan and Colin Batchelor

Argumentative Zoning (AZ) is an analysis of the argumentative and rhetorical structure of a scientific paper. It has been shown to be reliably recognisable by independent human coders, and has proven useful for various information access tasks. Annotation experiments have however so far been restricted to one discipline, computational linguistics (CL). Here, we present a more informative AZ scheme with 15 categories in place of the original 7, and show that it can be applied to the life sciences as well as to CL. We use a domain expert to encode basic knowledge about the subject (such as terminology and domain specific rules for individual categories) as part of the annotation guidelines. Our results show that non-expert human coders can then use these guidelines to reliably annotate this scheme in two domains, chemistry and computational linguistics.

Character-level Analysis of Semi-Structured Documents for Set Expansion

Richard C. Wang and William W. Cohen

Set expansion refers to expanding a partial set of “seed” objects into a more complete set. One system that does set expansion is SEAL (Set Expander for Any Language), which expands entities automatically by utilizing resources from the Web in a language-independent fashion. In this paper, we illustrated in detail the construction of character-level wrappers for set expansion implemented in SEAL. We also evaluated several kinds of wrappers for set expansion and showed that character-based wrappers perform better than HTML-based wrappers. In addition, we demonstrated a technique that extends SEAL to learn binary relational concepts (e.g., “x is the mayor of the city y”) from only two seeds. We also show that the extended SEAL has good performance on our evaluation datasets, which includes English and Chinese, thus demonstrating language-independence.

Classifying Relations for Biomedical Named Entity Disambiguation

Xinglong Wang, Jun'ichi Tsujii and Sophia Ananiadou

Linking an ambiguous mention of a named entity in text to an unambiguous identifier in a standard database is an important task for information extraction. Many conventional approaches treat named entity disambiguation as a supervised classification task. However, the availability of training data is often very limited, and the available data sets tend to be imbalanced and, in some cases, heterogeneous. We propose a new method that distinguishes a named entity by finding the informative keywords in its surrounding context, and then trains a model to predict whether each keyword indicates the semantic class of the entity. While maintaining a comparable performance to supervised classification, this method avoids using expensive manually annotated data for each new domain, and hence achieves better portability.

Domain adaptive bootstrapping for named entity recognition

Dan Wu, Wee Sun Lee, Nan Ye and Hai Leong Chieu

Bootstrapping is the process of improving the performance of a trained classifier by iteratively adding data that is labeled by the classifier itself to the training set, and retraining the classifier. It is often used in situations where labeled training data is scarce but unlabeled data is abundant. In this paper, we consider the problem of domain adaptation: the situation where training data may not be scarce, but belongs to a different domain from the target application domain. As the distribution of unlabeled data is different from the training data, standard bootstrapping often has difficulty selecting informative data to add to the training set. We propose an effective domain adaptive bootstrapping algorithm that selects unlabeled target domain data that are informative about the target domain and easy to automatically label correctly. We call these instances bridges, as they are used to bridge the source domain to the target domain. We show that the method outperforms supervised, transductive and bootstrapping algorithms on the named entity recognition task.

Phrase Dependency Parsing for Opinion Mining

Yuanbin Wu, Qi Zhang, Xuangjing Huang and Lide Wu

In this paper, we present a novel approach for mining opinions from product reviews, where it converts opinion mining task to identify product features, expressions of opinions and relations between them. By taking advantage of the observation that a lot of product features are phrases, a concept of phrase dependency parsing is introduced, which extends traditional dependency parsing to phrase level. This concept is then implemented for extracting relations between product features and expressions of opinions. Experimental evaluations show that the mining task can benefit from phrase dependency parsing.

Polynomial to Linear: Efficient Classification with Conjunctive Features

Naoki Yoshinaga and Masaru Kitsuregawa

This paper proposes a method that speeds up a classifier trained with many conjunctive features: combinations of (primitive) features. The key idea is to precompute as partial results the weights of primitive feature vectors that appear frequently in the target NLP task. A trie compactly stores the primitive feature vectors with their weights, and it enables the classifier to find for a given feature vector its longest prefix feature vector whose weight has already been computed. Experimental results for a Japanese dependency parsing task show that our method speeded up the SVM and LLM classifiers of the parsers, which achieved accuracy of 90.84/90.71%, by a factor of 10.7/11.6.

K-Best Combination of Syntactic Parsers

Hui Zhang, Min Zhang, Chew Lim Tan and Haizhou Li

In this paper, we propose a linear model-based general framework to combine k-best parse outputs from multiple parsers. The proposed framework leverages on the strengths of pre-vious system combination and re-ranking techniques in parsing by integrating them into a linear model. As a result, it is able to fully utilize both the logarithm of the probability of each k-best parse tree from each individual parser and any additional useful features. For feature weight tuning, we compare the simu-lated-annealing algorithm and the perceptron algorithm. Our experiments are carried out on both the Chinese and English Penn Treebank syntactic parsing task by combining two state-of-the-art parsing models, a head-driven lexi-calized model and a latent-annotation-based un-lexicalized model. Experimental results show that our F-Scores of 85.45 on Chinese and 92.62 on English outperform the previ-ously best-reported systems by 1.21 and 0.52, respectively.

Chinese Novelty Mining

Yi Zhang and Flora S. Tsai

Automated mining of novel documents or sentences from chronologically ordered documents or sentences is an open challenge in text mining. In this paper, we describe the preprocessing techniques for detecting novel Chinese text and discuss the influence of different Part of Speech (POS) filtering rules on the detection performance. Experimental results on APWSJ and TREC 2004 Novelty Track data show that the Chinese novelty mining performance is quite different when choosing two dissimilar POS filtering rules. Thus, the selection of words to represent Chinese text is of vital importance to the success of the Chinese novelty mining. Moreover, we compare the Chinese novelty mining performance with that of English and investigate the impact of preprocessing steps on detecting novel Chinese text, which will be very helpful for developing a Chinese novelty mining system.

Latent Document Re-Ranking

Dong Zhou and Vincent Wade

The problem of re-ranking initial retrieval results exploring the intrinsic structure of documents is widely researched in information retrieval (IR) and has attracted a considerable amount of time and study. However, one of the drawbacks is that those algorithms treat queries and documents separately. Furthermore, most of the approaches are predominantly built upon graph-based methods, which may ignore some hidden information among the retrieval set. This paper proposes a novel document re-ranking method based on Latent Dirichlet Allocation (LDA) which exploits the implicit structure of the documents with respect to original queries. Rather than relying on graph-based techniques to identify the internal structure, the approach tries to find the latent structure of “topics” or “concepts” in the initial retrieval set. Then we compute the distance between queries and initial retrieval results based on latent semantic information deduced. Empirical results demonstrate that the method can comfortably achieve significant improvement over various baseline systems.

Oral Presentations – Friday, 7 August

Multi-Word Expression Identification Using Sentence Surface Features

Ram Boukobza and Ari Rappoport

Much NLP research on Multi-Word Expressions (MWEs) focuses on the discovery of new expressions, as opposed to the identification in texts of known expressions. However, MWE identification is not trivial because many expressions allow variation in form and differ in the range of variations they allow. We show that simple rule-based baselines do not perform identification satisfactorily, and present a supervised learning method for identification that uses sentence surface features based on expressions' canonical form. To evaluate the method, we have annotated 3350 sentences from the British National Corpus, containing potential uses of 24 verbal MWEs. The method achieves an F-score of 94.86%, compared with 80.70% for the leading rule-based baseline. Our method is easily applicable to any expression type. Experiments in previous research have been limited to the compositional/non-compositional distinction, while we also test on sentences in which the words comprising the MWE appear but not as an expression.

Acquiring Translation Equivalences of Multiword Expressions by Normalized Correlation Frequencies

Ming-Hong Bai, Jia-Ming You, Keh-Jiann Chen and Jason S. Chang

In this paper, we present an algorithm for extracting translations of any given multiword expression from parallel corpora. Given a multiword expression to be translated, the method involves extracting a short list of target candidate words from parallel corpora based on scores of normalized frequency, generating possible translations and filtering out common subsequences, and selecting the top-n possible translations using the Dice coefficient. Experiments show that our approach outperforms the word alignment-based and other naive association-based methods. We also demonstrate that adopting the extracted translations can significantly improve the performance of the Moses machine translation system.

Collocation Extraction Using Monolingual Word Alignment Method

Zhanyi Liu, Haifeng Wang, Hua Wu and Sheng Li

Statistical bilingual word alignment has been well studied in the context of machine translation. This paper adapts the bilingual word alignment algorithm to monolingual scenario to extract collocations from monolingual corpus. The monolingual corpus is first replicated to generate a parallel corpus, where each sentence pair consists of two identical sentences in the same language. Then the monolingual word alignment algorithm is employed to align the potentially collocated words in the monolingual sentences. Finally the aligned word pairs are ranked according to refined alignment probabilities and those with higher scores are extracted as collocations. We conducted experiments using Chinese and English corpora individually. Compared with previous approaches, which use association measures to extract collocations from the co-occurring word pairs within a given window, our method achieves higher precision and recall. According to human evaluation in terms of precision, our method achieves absolute improvements of 27.9% on the Chinese corpus and 23.6% on the English corpus, respectively. Especially, we can extract collocations with longer spans, achieving a high precision of 69% on the long-span (>6) Chinese collocations.

Multi-Class Confidence Weighted Algorithms

Koby Crammer, Mark Dredze and Alex Kulesza

The recently introduced online confidence-weighted (CW) learning algorithm for binary classification performs well on many binary NLP tasks. However, for multi-class problems CW learning updates and inference cannot be computed analytically or solved as convex optimization problems as they are in the binary case. We derive learning algorithms for the multi-class CW setting and provide extensive evaluation using nine NLP datasets, including three derived from the recently released New York Times corpus. Our best algorithm outperforms state-of-the-art online and batch methods on eight of the nine tasks. We also show that the confidence information maintained during learning yields useful probabilistic information at test time.

Model Adaptation via Model Interpolation and Boosting for Web Search Ranking

Jianfeng Gao, Qiang Wu, Chris Burges, Krysta Svore, Yi Su, Nazan Khan, Shalin Shah and Hongyan Zhou

This paper explores two classes of model adaptation methods for Web search ranking: Model Interpolation and error-driven learning approaches based on a boosting algorithm. The results show that model interpolation, though simple, achieves the best results on all the open test sets where the test data is very different from the training data. The tree-based boosting algorithm achieves the best performance on most of the closed test sets where the test data and the training data are similar, but its performance drops significantly on the open test sets due to the instability of trees. Several methods are explored to improve the robustness of the algorithm, with limited success.

A Structural Support Vector Method for Extracting Contexts and Answers of Questions from Online Forums

Wen-Yun Yang, Yunbo Cao and Chin-Yew Lin

This paper addresses the issue of extracting contexts and answers of questions from post discussion of online forums. We propose a novel and unified model by customizing the structural Support Vector Machine method. Our customization has several attractive properties: (1) it gives a comprehensive graphical representation of thread discussion. (2) It designs special inference algorithms instead of general-purpose ones. (3) It can be readily extended to different task preferences by varying loss functions. Experimental results on a real data set show that our methods are both promising and flexible.

Mining Search Engine Clickthrough Log for Matching N-gram Features

Huihsin Tseng, Longbin Chen, Fan Li, Ziming Zhuang, Lei Duan and Belle Tseng

User clicks on a URL in response to a query are extremely useful predictors of the URL's relevance to that query. Exact match click features tend to suffer from severe data sparsity issues in web ranking. Such sparsity is particularly pronounced for new URLs or long queries where each distinct query-url pair will rarely occur. To remedy this, we present a set of straightforward yet informative query-url n-gram features that allows for generalization of limited user click data to large amounts of unseen query-url pairs. The method is motivated by techniques leveraged in the NLP community for dealing with unseen words. We find that there are interesting regularities across queries and their preferred destination URLs; for example, queries containing "form" tend to lead to clicks on URLs containing "pdf". We evaluate our set of new query-url features on a web search ranking task and obtain improve-ments that are statistically significant at a p-value < 0.0001 level over a strong baseline with exact match clickthrough features.

The role of named entities in Web People Search

Javier Artiles, Enrique Amigó and Julio Gonzalo

The ambiguity of person names in the Web has become a new area of interest for NLP researchers. This challenging problem has been formulated as the task of clustering Web search results (returned in response to a person name query) according to the individual they mention. In this paper we compare the coverage, reliability and independence of a number of features that are potential information sources for this clustering task, paying special attention to the role of named entities in the texts to be clustered. Although named entities are used in most approaches, our results show that, independently of the Machine Learning or Clustering algorithm used, named entity recognition and classification per se only make a small contribution to solve the problem.

Investigation of Question Classifier in Question Answering

Zhiheng Huang, Marcus Thint and Asli Celikyilmaz

In this paper, we investigate how an accurate question classifier contributes to a question answering system. We first propose a Maximum Entropy (ME) based question classifier which makes use of head word features and their WordNet hypernyms. We show that our question classifier can achieve the state of the art performance in the standard UIUC question dataset. We then investigate quantitatively the contribution of this question classifier to a feature driven question answering system. With our accurate question classifier and some standard question answer features, our question answering system performs close to the state of the art using TREC corpus.

An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing

Jun Suzuki, Hideki Isozaki, Xavier Carreras and Michael Collins

This paper describes an empirical study of high-performance dependency parsers based on a semi-supervised learning approach. We describe an extension of semi-supervised structured conditional models (SS-SCMs) to the dependency parsing problem, whose framework is originally proposed in (Suzuki and Isozaki, 2008). Moreover, we introduce two extensions related to dependency parsing: The first extension is to combine SS-SCMs with another semi-supervised approach, described in (Koo et al., 2008). The second extension is to apply the approach to second-order parsing models, such as those described in (Carreras, 2007), using a two-stage semi-supervised learning approach. We demonstrate the effectiveness of our proposed methods on dependency parsing experiments using two widely used test collections: the Penn Treebank for English, and the Prague Dependency Treebank for Czech. Our best results on test data in the above datasets achieve 93.79% parent-prediction accuracy for English, and 88.05% for Czech.

Statistical Bistratal Dependency Parsing

Richard Johansson

We present an inexact search algorithm for the problem of predicting a two-layered dependency graph. The algorithm is based on a k-best version of the standard cubic-time search algorithm for projective dependency parsing, which is used as the backbone of a beam search procedure. This allows us to handle the complex nonlocal feature dependencies occurring in bistratal parsing if we model the interdependency between the two layers. We apply the algorithm to the syntactic–semantic dependency parsing task of the CoNLL-2008 Shared Task, and we obtain a competitive result equal to the highest published for a system that jointly learns syntactic and semantic structure.

Improving Dependency Parsing with Subtrees from Auto-Parsed Data

Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto and Kentaro Torisawa

This paper presents a simple and effective approach to improve dependency parsing by using subtrees from auto-parsed data. First, we use a baseline parser to parse large-scale unannotated data. Then we extract subtrees from dependency parse trees in the auto-parsed data. Finally, we construct new subtree-based features for parsing algorithms. To demonstrate the effectiveness of our proposed approach, we present the experimental results on the English Penn Treebank and the Chinese Penn Treebank. These results show that our approach significantly outperforms baseline systems. And, it achieves the best accuracy for the Chinese data and an accuracy which is competitive with the best known systems for the English data.

Topic-wise, Sentiment-wise, or Otherwise? Identifying the Hidden Dimension for Unsupervised Text Classification

Sajib Dasgupta and Vincent Ng

While traditional work on text clustering has largely focused on grouping documents by topic, it is conceivable that a user may want to cluster the documents along other dimensions, such as the author's mood, gender, age, or sentiment. Without knowing the user's intention, a clustering algorithm will only group the documents along the most prominent dimension, which may not be the one the user desires. To address this problem, we propose a novel way of incorporating user feedback into a clustering algorithm, which allows a user to easily specify the dimension along which she wants the data points to be clustered via inspecting only a small number of words. This distinguishes our method from existing ones, which typically require a large amount of effort on the part of humans in the form of document annotation or interactive construction of the feature space. We demonstrate the viability of our approach on several challenging sentiment datasets.

Adapting a Polarity Lexicon using Integer Linear Programming for Domain-Specific Sentiment Classification

Yejin Choi and Claire Cardie

Polarity lexicons have been a valuable resource for sentiment analysis and opinion mining. There are a number of such lexical resources available, but it is often suboptimal to use them as is, because general purpose lexical resources do not reflect domain-specific lexical usage. In this paper, we propose a novel method based on integer linear programming that can adapt an existing lexicon into a new one to reflect the characteristics of the data more directly. In particular, our method collectively considers the relations among words and opinion expressions to derive the most likely polarity of each lexical item (positive, neutral, negative, or negator) for the given domain. Experimental results show that our lexicon adaptation technique improves the performance of fine-grained polarity classification.

Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus

Saif Mohammad, Cody Dunne and Bonnie Dorr

Sentiment analysis often relies on a semantic orientation lexicon of positive and negative words. A number of approaches have been proposed for creating such lexicons, but they tend to be computationally expensive, and usually rely on significant manual annotation and large corpora. Most of these methods use WordNet. In contrast, we propose a simple approach to generate a high-coverage semantic orientation lexicon, which includes both individual words and multi-word expressions, using only a Roget-like thesaurus and a handful of affixes. Further, the lexicon has properties that support the Polyanna Hypothesis. Using the General Inquirer as gold standard, we show that our lexicon has 14 percentage points more correct entries than the leading WordNet-based high-coverage lexicon (SentiWordNet). In an extrinsic evaluation, we obtain significantly higher performance in determining phrase polarity using our thesaurus-based lexicon than with any other. Additionally, we explore the use of visualization techniques to gain insight into the our algorithm beyond the evaluations mentioned above.

Matching Reviews to Objects using a Language Model

Nilesh Dalvi, Ravi Kumar, Bo Pang and Andrew Tomkins

We develop a general method to match unstructured text reviews to a structured list of objects. For this, we propose a language model for generating reviews that incorporates a description of objects and a generic review language model. This mixture model gives us a principled method to find, given a review, the object most likely to be the topic of the review. Extensive experiments and analysis on reviews from Yelp show that our language model-based method vastly outperforms traditional tf-idf-based methods.

EEG responds to conceptual stimuli and corpus semantics

Brian Murphy, Marco Baroni and Massimo Poesio

Mitchell et al. (2008) demonstrated that corpus-extracted models of semantic knowledge can predict neural activation patterns recorded using fMRI. This could be a very powerful technique for evaluating conceptual models extracted from corpora; however, fMRI is expensive and imposes strong constraints on data collection. Following on experiments that demonstrated that EEG activation patterns encode enough information to discriminate broad conceptual categories, we show that corpus-based semantic representations can predict EEG activation patterns with significant accuracy, and we evaluate the relative performance of different corpus-models on this task.

A Comparison ofWindowless and Window-Based Computational Association Measures as Predictors of Syntagmatic Human Associations

Justin Washtell and Katja Markert

Distance-based (windowless) word assocation measures have only very recently appeared in the NLP literature and their performance compared to existing windowed or frequency-based measures is largely unknown. We conduct a large-scale empirical comparison of a variety of distance-based and frequency-based measures for the reproduction of syntagmatic human assocation norms. Overall, our results show an improvement in the predictive power of windowless over windowed measures. This provides support to some of the previously published theoretical advantages and makes windowless approaches a promising avenue to explore further. This study also serves as a first comparison of windowed methods across numerous human association datasets. During this comparison we also introduce some novel variations of window-based measures which perform as well as or better in the human association norm task than established measures.

Improving Verb Clustering with Automatically Acquired Selectional Preferences

Lin Sun and Anna Korhonen

In previous research in automatic verb classification, syntactic features have proved the most useful features, although manual classifications rely heavily on semantic features. We show, in contrast with previous work, that considerable additional improvement can be obtained by using semantic features in automatic classification: verb selectional preferences acquired from corpus data using a fully unsupervised method. We report these promising results using a new framework for verb clustering which incorporates a recent subcategorization acquisition system, rich syntactic-semantic feature sets, and a variation of spectral clustering which performs particularly well in high dimensional feature space.

Improving Web Search Relevance with Semantic Features

Yumao Lu, Fuchun Peng, Gilad Mishne, Xing Wei and Benoit Dumoulin

Most existing information retrieval (IR) systems do not take much advantage of natural language processing (NLP) techniques due to the complexity and limited observed effectiveness of applying NLP to IR. In this paper, we demonstrate that substantial gains can be obtained over a strong baseline using NLP techniques, if properly handled. We propose a framework for deriving semantic text matching features from named entities identified in Web queries; we then utilize these features in a supervised machine-learned ranking approach, applying a set of emerging machine learning techniques. Our approach is especially useful for queries that contain multiple types of concepts. Comparing to a major commercial Web search engine, we observe a substantial 4% DCG5 gain over the affected queries.

Can Chinese Phonemes Improve Machine Transliteration?: A Comparative Study of English-to-Chinese Transliteration Models

Jong-Hoon Oh, Kiyotaka Uchimoto and Kentaro Torisawa

Inspired by the success of English grapheme-to-phoneme research in speech synthesis, many researchers have proposed phoneme-based English-to-Chinese transliteration models. However, such approaches have severely suffered from the errors in Chinese phoneme-to-grapheme conversion. To address this issue, we propose a new English-to-Chinese transliteration model and make systematic comparisons with the conventional models. Our proposed model relies on the joint use of Chinese phonemes and their corresponding English graphemes and phonemes. Experiments showed that Chinese phonemes in our proposed model can contribute to the performance improvement in English-to-Chinese transliteration.

Unsupervised morphological segmentation and clustering with document boundaries

Taesun Moon, Katrin Erk and Jason Baldridge

Many approaches to unsupervised morphology acquisition incorporate the frequency of character sequences with respect to each other to identify word stems and affixes. This typically involves heuristic search procedures and calibrating multiple arbitrary thresholds. We present a simple approach that uses no thresholds other than those involved in standard application of chi-square significance testing. A key part of our approach is using document boundaries to constrain generation of candidate stems and affixes and clustering morphological variants of a given word stem. We evaluate our model on English and the Mayan language Uspanteko; it compares favorably to two benchmark systems which use considerably more complex strategies and rely more on experimentally chosen threshold values.

The infinite HMM for unsupervised PoS tagging

Jurgen Van Gael, Andreas Vlachos and Zoubin Ghahramani

We extend previous work on fully unsupervised part-of-speech tagging. Using a non-parametric version of the HMM, called the infinite HMM (iHMM), we address the problem of choosing the number of hidden states in unsupervised Markov models for PoS tagging. We experiment with two non-parametric priors, the Dirichlet and Pitman-Yor processes, on the Wall Street Journal dataset using a parallelized implementation of an iHMM inference algorithm. We evaluate the results with a variety of clustering evaluation metrics and achieve equivalent or better performances than previously reported. Building on this promising result we evaluate the output of the unsupervised PoS tagger as a direct replacement for the output of a fully supervised PoS tagger for the task of shallow parsing and compare the two evaluations.

A Simple Unsupervised Learner for POS Disambiguation Rules Given Only a Minimal Lexicon.

Qiuye Zhao and Mitch Marcus

We propose a new model for unsupervised POS tagging based on linguistic distinctions between open and closed-class items. Exploiting notions from current linguistic theory, the system uses far less information than previous systems, far simpler computational methods, and far sparser descriptions in learning contexts. By applying simple language acquisition techniques based on counting, the system is given the closed-class lexicon, acquires a large open-class lexicon and then acquires disambiguation rules for both. This system achieves a 20% error reduction for POS tagging over state-of-the-art unsupervised systems tested under the same conditions, and achieves comparable accuracy when trained with much less prior information.

Tree Kernel-based SVM with Structured Syntactic Knowledge for BTG-based Phrase Reordering

Min Zhang and Haizhou Li

Structured syntactic knowledge is important for phrase reordering. This paper proposes using convolution tree kernel over source parse tree to model structured syntactic knowledge for BTG-based phrase reordering in the context of statistical machine translation. Our study reveals that the structured syntactic features over the source phrases are very effective for BTG constraint-based phrase reordering and those features can be well captured by the tree kernel. We further combine the structured features and other commonly-used linear features into a composite kernel. Experimental results on the NIST MT-2005 Chinese-English translation tasks show that our proposed phrase reordering model statistically significantly outperforms the baseline methods.

Discriminative Corpus Weight Estimation for Machine Translation

Spyros Matsoukas, Antti-Veikko I. Rosti and Bing Zhang

Current statistical machine translation (SMT) systems are trained on sentence-aligned and word-aligned parallel text collected from various sources. Translation model parameters are estimated from the word alignments, and the quality of the translations on a given test set depends on the parameter estimates. There are at least two factors affecting the parameter estimation: domain match and training data quality. This paper describes a novel approach for automatically detecting and down-weighing certain parts of the training corpus by assigning a weight to each sentence in the training bitext so as to optimize a discriminative objective function on a designated tuning set. This way, the proposed method can limit the negative effects of low quality training data, and can adapt the translation model to the domain of interest. It is shown that such discriminative corpus weights can provide significant improvements in Arabic-English translation on various conditions, using a state-of-the-art SMT system.

Unsupervised Tokenization for Machine Translation

Tagyoung Chung and Daniel Gildea

Training a statistical machine translation starts with tokenizing a parallel corpus. Some languages such as Chinese do not incorporate spacing in their writing system, which creates a challenge for tokenization. Moreover, morphologically rich languages such as Korean present an even bigger challenge, since optimal token boundaries for machine translation in these languages are often unclear. Both rule-based solutions and statistical solutions are currently used. In this paper, we present unsupervised methods to solve tokenization problem. Our methods incorporate information available from parallel corpus to determine a good tokenization for machine translation.

Synchronous Tree Adjoining Machine Translation

Steve DeNeefe and Kevin Knight

Tree Adjoining Grammars have well-known advantages, but are typically considered too difficult for practical systems. We demonstrate that, when done right, adjoining improves translation quality without becoming computationally intractable. Using adjoining to model optionality allows general translation patterns to be learned without the clutter of endless variations of optional material. The appropriate modifiers can later be spliced in as needed. In this paper, we describe a novel method for learning a type of Synchronous Tree Adjoining Grammar and associated probabilities from aligned tree/string training data. We introduce a method of converting these grammars to a weakly equivalent tree transducer for decoding. Finally, we show that adjoining results in an end-to-end improvement of +0.8 BLEU over a baseline statistical syntax-based MT model on a large-scale Arabic/English MT task.

Word Buffering Models for Improved Speech Repair Parsing

Tim Miller

This paper describes a time-series model for parsing transcribed speech containing disfluencies. This model differs from previous parsers in its explicit modeling of a buffer of recent words, which allows it to recognize repairs more easily due to the frequent overlap in words between errors and their repairs. The parser implementing this model is evaluated on the standard Switchboard transcribed speech parsing task for overall parsing accuracy and edited word detection.

Less is More: Significance-Based N-gram Selection for Smaller, Better Language Models

Robert C. Moore and Chris Quirk

The recent availability of large corpora for training N-gram language models has shown the utility of models of higher order than just trigrams. In this paper, we investigate methods to control the increase in model size resulting from applying standard methods at higher orders. We introduce significance-based N-gram selection, which not only reduces model size, but also improves perplexity for several smoothing methods, including Katz backoff and absolute discounting. We also show that, when combined with a new smoothing method and a novel variant of weighted-difference pruning, our selection method performs better in the trade-off between model size and perplexity than the best pruning method we found for modified Kneser-Ney smoothing.

Stream-based Randomised Language Models for SMT

Abby Levenberg and Miles Osborne

Randomised techniques allow very big language models to be represented succinctly. However, being batch-based they are unsuitable for modelling an unbounded stream of language whilst maintaining a constant error rate. We present a novel randomised language model which uses an online perfect hash function to efficiently deal with unbounded text streams. Translation experiments over a text stream show that our online randomised model matches the performance of batch-based LMs without incurring the computational overhead associated with full retraining. This opens up the possibility of randomised language models which continuously adapt to the massive volumes of texts published on the Web each day.

Integrating sentence- and word-level error identification for disfluency correction

Erin Fitzgerald, Frederick Jelinek and Keith Hall

While speaking spontaneously, speakers often make errors such as self-correction or false starts which interfere with the successful application of natural language processing techniques like summarization and machine translation to this data. There is active work on reconstructing this errorful data into a clean and fluent transcript by identifying and removing these simple errors. Previous research has approximated the potential benefit of conducting word-level reconstruction of simple errors only on those sentences known to have errors. In this work, we explore new approaches for automatically identifying speaker construction errors on the utterance level, and quantify the impact that this initial step has on word- and sentence-level reconstruction accuracy.

Estimating Semantic Distance Using Soft Semantic Constraints in Knowledge-Source – Corpus Hybrid Models

Yuval Marton, Saif Mohammad and Philip Resnik

Strictly corpus-based measures of semantic distance conflate co-occurrence information pertaining to the many possible senses of target words. We propose a corpusthesaurus hybrid method that uses soft constraints to generate word-sense- aware distributional profiles (DPs) from coarser “concept DPs” (derived from a Roget-like thesaurus) and sense-unaware traditional word DPs (derived from raw text). Although it uses a knowledge source, the method is not vocabulary-limited: if the target word is not in the thesaurus, the method falls back gracefully on the word’s co-occurrence information. This allows the method to access valuable information encoded in a lexical resource, such as a thesaurus, while still being able to effectively handle domain-specific terms and named entities. Experiments on word-pair ranking by semantic distance show the new hybrid method to be superior to others.

Recognizing Textual Relatedness with Predicate-Argument Structures

Rui Wang and Yi Zhang

In this paper, we first compare several strategies to handle the newly proposed three-way Recognizing Textual Entailment (RTE) task. Then we define a new measurement for a pair of texts, called Textual Relatedness, which is a weaker concept than semantic similarity or paraphrase. We show that an alignment model based on the predicate-argument structures using this measurement can help an RTE system to recognize the Unknown cases at the first stage, and contribute to the improvement of the overall performance in the RTE task. In addition, several heterogeneous lexical resources are tested, and different contributions from them are observed.

Learning Term-weighting Functions for Similarity Measures

Wen-tau Yih

Measuring the similarity between two texts is a fundamental problem in many NLP and IR applications. Among the existing approaches, the cosine measure of the term vectors representing the original texts has been widely used, where the score of each term is often determined by a TFIDF formula. Despite its simplicity, the quality of such cosine similarity measure is usually domain dependent and decided by the choice of the term-weighting function. In this paper, we propose a novel framework that learns the term-weighting function. Given the labeled pairs of texts as training data, the learning procedure tunes the model parameters by minimizing the specified loss function of the similarity score. Compared to traditional TFIDF term-weighting schemes, our approach shows a significant improvement on tasks such as judging the quality of query suggestions and filtering irrelevant ads for online advertising.

A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web

Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka

Semantic similarity is a central concept that extends across numerous fields such as artificial intelligence, natural language processing, cognitive science and psychology. Accurate measurement of semantic similarity between words is essential for various tasks such as, document clustering, information retrieval, and synonym extraction. We propose a novel model of semantic similarity using the semantic relations that exist among words. Given two words, first, we represent the semantic relations that hold between those words using automatically extracted lexical pattern clusters. Next, the semantic similarity between the two words is computed using a Mahalanobis distance measure. We compare the proposed similarity measure against previously proposed semantic similarity measures on Miller-Charles benchmark dataset and WordSimilarity-353 collection. The proposed method outperforms all existing web-based semantic similarity measures, achieving a Pearson correlation coefficient of 0.867 on the Millet-Charles dataset.

Unbounded Dependency Recovery for Parser Evaluation

Laura Rimell, Stephen Clark and Mark Steedman

This paper introduces a new parser evaluation corpus containing around 700 sentences annotated with unbounded dependencies, from seven different grammatical constructions. We run a series of off-the-shelf parsers on the corpus to evaluate how well state-of-the-art parsing technology is able to recover such dependencies. The overall results range from 25% accuracy to 59%. These low scores call into question the validity of using Parseval scores as a general measure of parsing capability. We discuss the importance of parsers being able to recover unbounded dependencies, given their relatively low frequency in corpora. We also analyse the various errors made on these constructions by one of the more successful parsers.

Parser Adaptation and Projection with Quasi-Synchronous Grammar Features

David A. Smith and Jason Eisner

We connect two scenarios in structured learning: adapting a parser trained on one corpus to another annotation style, and projecting syntactic annotations from one language to another. We propose quasi-synchronous grammar (QG) features for these structured learning tasks. That is, we score an aligned pair of source and target trees based on local features of the trees and the alignment. Our quasi-synchronous model assigns positive probability to any alignment of any trees, in contrast to a synchronous grammar, which would insist on some form of structural parallelism. In monolingual dependency parser adaptation, we achieve high accuracy in translating among multiple annotation styles for the same sentence. On the more difficult problem of cross-lingual parser projection, we learn a dependency parser for a target language by using bilingual text, an English parser, and automatic word alignments. Our experiments show that unsupervised QG projection improves on parses trained using only high-precision projected annotations and far outperforms, by more than 35% absolute dependency accuracy, learning an unsupervised parser from raw target-language text alone. When a few target-language parse trees are available, projection gives a boost equivalent to doubling the number of target-language trees.

Self-Training PCFG Grammars with Latent Annotations Across Languages

Zhongqiang Huang and Mary Harper

We investigate the effectiveness of self-training PCFG grammars with latent annotations (PCFG-LA) for parsing languages with different amounts of labeled training data. Compared to Charniak's lexicalized parser, the PCFG-LA parser was more effectively adapted to a language for which parsing has been less well developed (i.e., Chinese) and benefited more from self-training. We show for the first time that self-training is able to significantly improve the performance of the PCFG-LA parser, a single generative parser, on both small and large amounts of labeled training data. Our approach achieves state-of-the-art parsing accuracies for a single parser on both English (91.5%) and Chinese (85.2%).

An Alternative to Head-Driven Approaches for Parsing a (Relatively) Free Word-Order Language

Reut Tsarfaty, Khalil Sima'an and Remko Scha

Applying statistical parsers developed for English to languages with freer word-order has turned out to be harder than expected. This paper investigates the adequacy of different statistical parsing models for dealing with a (relatively) free word-order language. We show that the recently proposed Relational-Realizational (RR) model consistently outperforms state-of-the-art Head-Driven (HD) models on the Hebrew Treebank. Our analysis reveals a weakness of HD models: their intrinsic focus on configurational information. We conclude that the form-function separation ingrained in RR models makes them better suited for parsing nonconfigurational phenomena.

Enhancement of Lexical Concepts Using Cross-lingual Web Mining

Dmitry Davidov and Ari Rappoport

Sets of lexical items sharing a significant aspect of their meaning (concepts) are fundamental in linguistics and NLP. Manual concept compilation is labor intensive, error prone and subjective. We present a web-based concept extension algorithm. Given a set of terms specifying a concept in some language, we translate them to a wide range of intermediate languages, disambiguate the translations using web counts, and discover additional concept terms using symmetric patterns. We then translate the discovered terms back into the original language, score them, and extend the original concept by adding back-translations having high scores. We evaluate our method in 3 source languages and 45 intermediate languages, using both human judgments and WordNet. In all cases, our cross-lingual algorithm significantly improves high quality concept extension.

Bilingual dictionary generation for low-resourced language pairs

István Varga and Shoichi Yokoyama

Bilingual dictionaries are vital resources in many areas of natural language processing. Numerous methods of machine translation require bilingual dictionaries with large coverage, but less-frequent language pairs rarely have any digitalized resources. Since the need for these resources is increasing, but the human resources are scarce for less represented languages, efficient automatized methods are needed. This paper introduces a fully automated, robust pivot language based bilingual dictionary generation method that uses the WordNet of the pivot language to build a new bilingual dictionary. We propose the usage of WordNet in order to increase accuracy; we also introduce a bidirectional selection method with a flexible threshold to maximize recall. Our evaluations showed 79% accuracy and 51% weighted recall, outperforming representative pivot language based methods. A dictionary generated with this method will still need manual post-editing, but the improved recall and precision decrease the work of human correctors.

Multilingual Spectral Clustering Using Document Similarity Propagation

Dani Yogatama and Kumiko Tanaka-Ishii

We present a novel approach for multilingual document clustering using only comparable corpora to achieve cross-lingual semantic interoperability. The method models document collections as weighted graph, and supervisory information is given as sets of must-linked constraints for documents in different languages. Recursive k-nearest neighbor similarity propagation is used to exploit the prior knowledge and merge two language spaces. Spectral method is applied to find the best cuts of the graph. Experimental results show that using limited supervisory information, our method achieves promising clustering results. Furthermore, since the method does not need any language dependent information in the process, our algorithm can be applied to languages in various alphabetical systems.

Polylingual Topic Models

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith and Andrew McCallum

Topic models are a useful tool for analyzing large text collections, but have previously been applied in only monolingual, or at most bilingual, contexts. Meanwhile, massive collections of interlinked documents in dozens of languages, such as Wikipedia, are now widely available, calling for tools that can characterize content in many languages. We introduce a polylingual topic model that discovers topics aligned across multiple languages. We explore the model's characteristics using two large corpora, each with over ten different languages, and demonstrate its usefulness in supporting machine translation and tracking topic trends across languages.

Using the Web for Language Independent Spellchecking and Autocorrection

Casey Whitelaw, Ben Hutchinson, Grace Y. Chung and Ged Ellis

We have designed, implemented and evaluated an end-to-end system spellchecking and autocorrection system that does not require any manually annotated training data. The World Wide Web is used as a large noisy corpus from which we infer knowledge about misspellings and word usage. This is used to build an error model and an n-gram language model. A small secondary set of news texts with artificially inserted misspellings are used to tune confidence classifiers. Because no manual annotation is required, our system can easily be instantiated for new languages. When evaluated on human typed data with real misspellings in English and German, our web-based systems outperform baselines which use candidate corrections based on hand-curated dictionaries. Our system achieves 3.8% total error rate in English. We show similar improvements in preliminary results on artificial data for Russian and Arabic.

Statistical Estimation of Word Acquisition with Application to Readability Prediction

Paul Kidwell, Guy Lebanon and Kevyn Collins-Thompson

Models of language learning play a central role in a wide range of applications: from psycholinguistic theories of how people acquire new word knowledge, to information systems that can automatically match content to users' reading ability. We present a novel statistical approach that can infer the distribution of a word's likely acquisition age automatically from authentic texts collected from the Web. We then show that combining these acquisition age distributions for all words in a document provides an effective semantic component for predicting reading difficulty of new texts. We also compare our automatically inferred acquisition ages with norms from existing oral studies, revealing interesting historical trends as well as differences between oral and written word acquisition processes.

Combining Collocations, Lexical and Encyclopedic Knowledge for Metonymy Resolution

Vivi Nastase and Michael Strube

This paper presents a supervised method for resolving metonymies. We enhance a commonly used feature set with features extracted based on collocation information from corpora, generalized using lexical and encyclopedic knowledge to determine the preferred sense of the potentially metonymic word using methods from unsupervised word sense disambiguation. The methodology developed addresses one issue related to metonymy resolution – the influence of local context. The method developed is applied to the metonymy resolution task from SemEval 2007. The results obtained, higher for the countries subtask, on a par for the companies subtask – compared to participating systems – confirm that lexical, encyclopedic and collocation information can be successfully combined for metonymy resolution.

Segmenting Email Message Text into Zones

Andrew Lampert, Robert Dale and Cécile Paris

In the early days of email, widely-used conventions for indicating quoted reply content and email signatures made it easy to segment email messages into their functional parts. Today, the explosion of different email formats and styles, coupled with the ad hoc ways in which people vary the structure and layout of their messages, means that simple techniques for identifying quoted replies that used to yield 95% accuracy now find less than 10% of such content. In this paper, we describe Zebra, an SVM-based system for segmenting the body text of email messages into nine zone types based on graphic, orthographic and lexical cues. Zebra performs this task with an accuracy of 87.01%; when the number of zones is abstracted to two or three zone classes, this increases to 93.60% and 91.53% respectively.

Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures

Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond and Asuka Sumida

This paper presents a new method of developing a large-scale hyponymy relation database by combining Wikipedia and other Web documents. We attach new words to the hyponymy database extracted from Wikipedia by using distributional similarity calculated from documents on the Web. For a given target word, our algorithm first finds k similar words from the Wikipedia database. Then, the hypernyms of these k similar words are assigned scores by considering the distributional similarities and hierarchical distances in the Wikipedia database. Finally, new hyponymy relations are output according to the scores. In this paper, we tested two distributional similarities. One is based on raw verb-noun dependencies (which we call RVD), and the other is based on a large-scale clustering of verb-noun dependencies (called CVD). Our method achieved an attachment accuracy of 91.0% for the top 10,000 relations, and an attachment accuracy of 74.5% for the top 100,000 relations when using CVD. This was a far better outcome compared to the other baseline approaches. Excluding the region that had very high scores, CVD was found to be more effective than RVD. We also confirmed that most relations extracted by our method cannot be extracted merely by applying the well-known lexico-syntactic patterns to Web documents.

Web-Scale Distributional Similarity and Entity Set Expansion

Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu and Vishnu Vyas

Computing the pairwise semantic similarity between all words on the Web is a computationally challenging task. Parallelization and optimizations are necessary. We propose a highly scalable implementation based on distributional similarity, implemented in the MapReduce framework and deployed over a 200 billion word crawl of the Web. The pairwise similarity between 500 million terms is computed in 50 hours using 200 quad-core nodes. We apply the learned similarity matrix to the task of automatic set expansion and present a large empirical study to quantify the effect on expansion performance of corpus size, corpus quality, seed composition and seed size. We make public an experimental testbed for set expansion analysis that includes a large collection of diverse entity sets extracted from Wikipedia.

Toward Completeness in Concept Extraction and Classification

Eduard Hovy, Zornitsa Kozareva and Ellen Riloff

Many algorithms extract terms from text together with some kind of taxonomic classification (is-a) link. However, the general approach exhibits serious shortcomings. Harvesting without focusing on a specific conceptual area may deliver large numbers of terms, but they are scattered over an immense concept space, making Recall judgments impossible. Regarding Precision, simply judging the correctness of terms and their individual classification links may provide high scores, but this doesn't help with the eventual assembly into a single coherent taxonomy. Furthermore, there is no correct and complete gold standard to measure against, and most work invents some ad hoc evaluation measure. We present an algorithm that is more precise and complete than previous ones for identifying from text just those concepts 'below' a given seed term, for subsequent organization into a taxonomy. Comparing the results to WordNet, we find that the algorithm misses terms, but also that it learns many new terms not in WordNet, and that it classifies them in ways acceptable to humans but different from WordNet.

Reading to Learn: Constructing Features from Semantic Abstracts

Jacob Eisenstein, James Clarke, Dan Goldwasser and Dan Roth

Machine learning offers a range of tools for training systems from data, but these methods are only as good as the underlying representation. This paper proposes to acquire representations for machine learning by reading text written to accommodate human learning. We propose a novel form of semantic analysis called reading to learn, where the goal is to obtain a high-level semantic abstract of multiple documents in a representation that facilitates learning. We obtain this abstract through a generative model that requires no labeled data, instead leveraging repetition across multiple documents. The semantic abstract is converted into a transformed feature space for learning, resulting in improved generalization on a relational learning task.

Supervised Models for Coreference Resolution

Altaf Rahman and Vincent Ng

Traditional learning-based coreference resolvers operate by training a mention-pair classifier for determining whether two mentions are coreferent or not. Two independent lines of recent research have attempted to improve these mention-pair classifiers, one by learning a mention-ranking model to rank preceding mentions for a given anaphor, and the other by training an entity-mention classifier to determine whether a preceding cluster is coreferent with a given mention. We propose a cluster-ranking approach to coreference resolution that combines the strengths of mention rankers and entity-mention models. We additionally show how our cluster-ranking framework naturally allows the detection of discourse-new entities to be learned jointly with coreference resolution. Experimental results on the ACE data sets demonstrate its superior performance to competing approaches.

Global Learning of Noun Phrase Anaphoricity in Coreference Resolution via Label Propagation

GuoDong Zhou and Fang Kong

Knowledge of noun phrase anaphoricity might be profitably exploited in coreference resolution to bypass the resolution of non-anaphoric noun phrases. However, it is surprising to notice that recent attempts to incorporate automatically acquired anaphoricity information into coreference resolution have been somewhat disappointing. This paper employs a global learning method in determining the anaphoricity of noun phrases via a label propagation algorithm to improve learning-based coreference resolution. In particular, two kinds of kernels, i.e. the feature-based RBF kernel and the convolution tree kernel, are employed to compute the anaphoricity similarity between two noun phrases. Experiments on the ACE 2003 corpus demonstrates the effectiveness of our method in anaphoricity determination of noun phrases and its application in learning-based coreference resolution.

Employing the Centering Theory in Pronoun Resolution from the Semantic Perspective

Fang Kong, GuoDong Zhou and Qiaoming Zhu

In this paper, we employ the centering theory in pronoun resolution from the semantic perspective. First, diverse semantic role features with regard to different predicates in a sentence are explored. Moreover, given a pronominal anaphor, its relative ranking among all the pronouns in a sentence, according to relevant semantic role information and its surface position, is incorporated. In particular, the use of both the semantic role features and the relative pronominal ranking feature in pronoun resolution is guided by extending the centering theory from the grammatical level to the semantic level in tracking the local discourse structure. Finally, detailed pronominal subcategory features are incorporated to enhance the discriminative power of both the semantic role features and the relative pronominal ranking feature. Experimental results on the ACE 2003 corpus show that the centering-motivated features contribute much to pronoun resolution.

Person Cross Document Coreference with Name Perplexity Estimates

Octavian Popescu

The Person Cross Document Coreference systems depend on the context for making decisions on the possible coreferences between person name mentions. The amount of context required is a parameter that varies from corpora to corpora, which makes it difficult for usual disambiguation methods. In this paper we show that the amount of context required can be dynamically controlled on the basis of the prior probabilities of coreference and we present a new statistical model for the computation of these probabilities. The experiment we carried on a news corpus proves that the prior probabilities of coreference are an impor-tant factor for maintaining a good balance be-tween precision and recall for cross document coreference systems.

Learning Linear Ordering Problems for Better Translation

Roy Tromble and Jason Eisner

We apply machine learning to the Linear Ordering Problem in order to learn sentence-specific reordering models for machine translation. We demonstrate that even when these models are used as a mere preprocessing step for German-English translation, they significantly outperform Moses' integrated lexicalized reordering model. Our models are trained on automatically aligned bitext. Their form is simple but novel. They assess, based on features of the input sentence, how strongly each pair of input word tokens w_i, w_j would like to reverse their relative order. Combining all these pairwise preferences to find the best global reordering is NP-hard. However, we present a non-trivial $O(n^3)$ algorithm, based on chart parsing, that at least finds the best reordering within a certain exponentially large neighborhood. We show how to iterate this reordering process within a local search algorithm, which we use in training.

Weighted Alignment Matrices for Statistical Machine Translation

Yang Liu, Tian Xia, Xinyan Xiao and Qun Liu

Current statistical machine translation systems usually extract rules from bilingual corpora annotated with 1-best alignments. They are prone to learn noisy rules due to alignment mistakes. We propose a new structure called weighted alignment matrix to encode all possible alignments for a parallel text compactly. The key idea is to assign a probability to each word pair to indicate how well they are aligned. We design new algorithms for extracting phrase pairs from weighted alignment matrices and estimating their probabilities. Our experiments on multiple language pairs show that using weighted matrices achieves consistent improvements over using n-best lists in significant less extraction time.

Sinuhe – Statistical Machine Translation using a Globally Trained Conditional Exponential Family Translation Model

Matti Kääriäinen

We present a new phrase-based conditional exponential family translation model for statistical machine translation. The model operates on a feature representation in which sentence level translations are represented by enumerating all the known phrase level translations that occur inside them. This makes the model a good match with the commonly used phrase extraction heuristics. The model's predictions are properly normalized probabilities. In addition, the model automatically takes into account information provided by phrase overlaps, and does not suffer from reference translation reachability problems. We have implemented an open source translation system Sinuhe based on the proposed translation model. Our experiments on Europarl and GigaFrEn corpora demonstrate that finding the unique MAP parameters for the model on large scale data is feasible with simple stochastic gradient methods. Sinuhe is fast and memory efficient, and the BLEU scores obtained by it are only slightly inferior to those of Moses.

Fast Translation Rule Matching for Syntax-based Statistical Machine Translation

Hui Zhang, Min Zhang, Haizhou Li and Chew Lim Tan

In a linguistically-motivated syntax-based translation system, the entire translation process is normally carried out in two steps, translation rule matching and target sentence decoding using the matched rules. Both steps are very time-consuming due to the exponential number of translation rules, the exhaustive search in translation rule matching and the complex nature of the translation task itself. In this paper, we propose a hyper-tree-based fast algorithm for translation rule matching. Experimental results on the NIST MT-2003 Chinese-English translation task show that our algorithm is at least 19 times faster in rule matching and is able to help to save 57% of overall translation time over previous methods when using large fragment translation rules.

Index

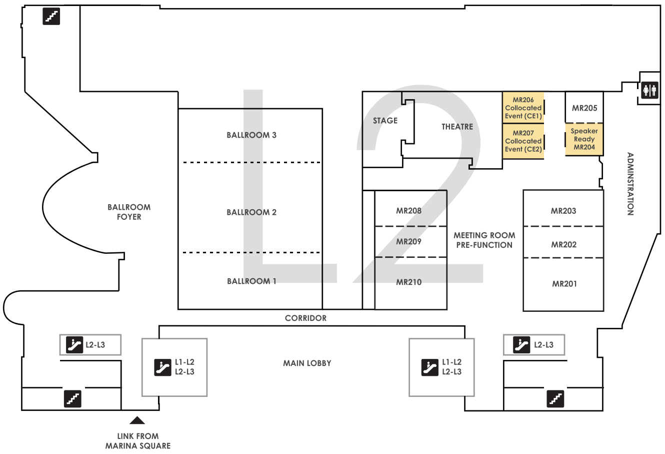
13

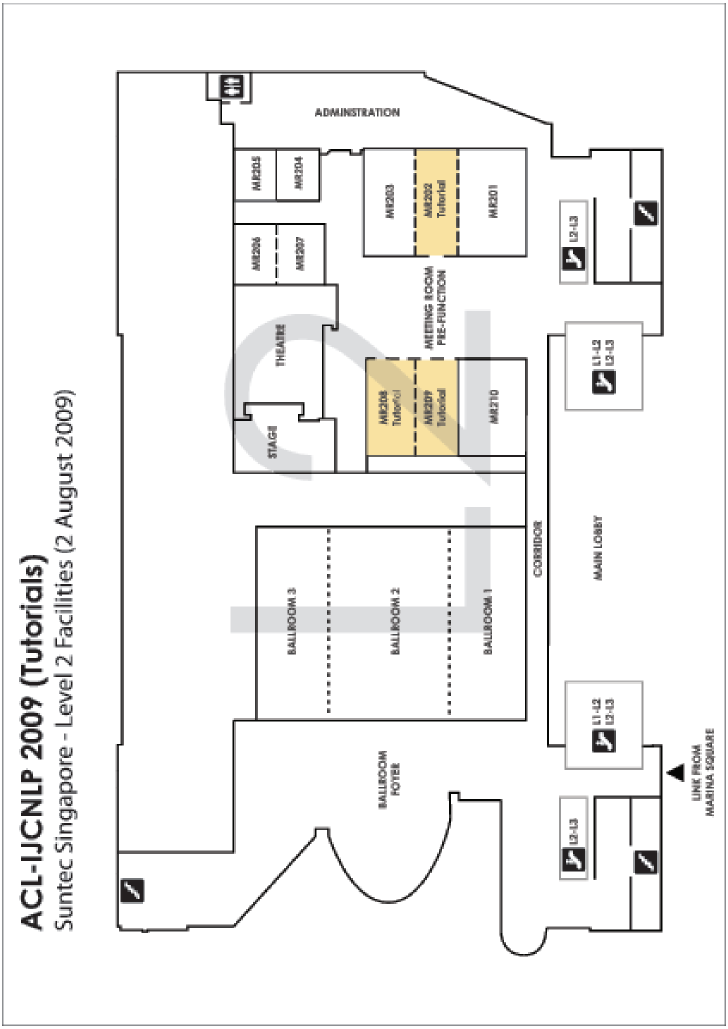
Maps

Pre-Conference

ACL-IJCNLP 2009 (Workshop and Co-Located Events)

Suntec Singapore - Level 2 Facilities (2 August 2009)



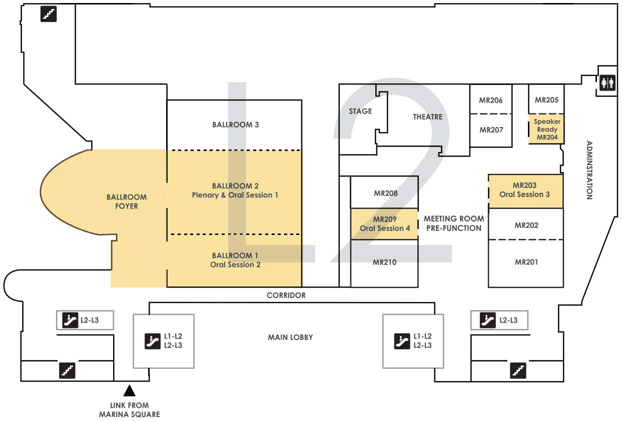


ACL-IJCNLP 2009 (Tutorials)

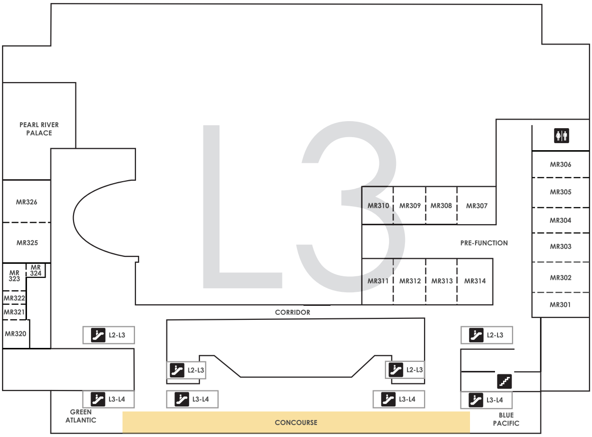
Suntec Singapore - Level 2 Facilities (2 August 2009)

ACL-IJCNLP Conference

ACL-IJCNLP 2009
(Main Conference with Registration/Exhibition/Demos/Posters)
Suntec Singapore - Level 2 Facilities (3-5 August 2009)

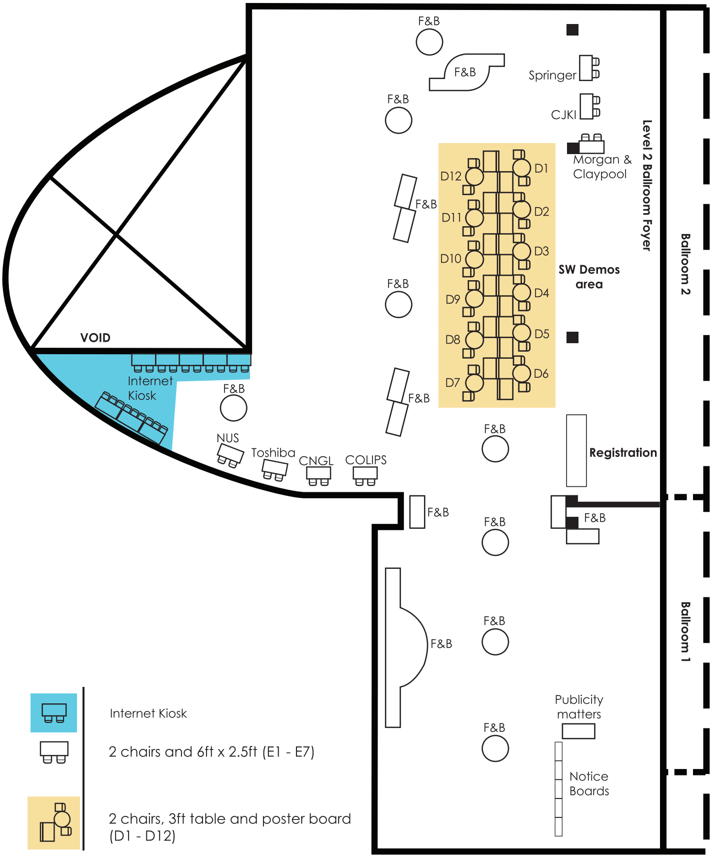


ACL-IJCNLP 2009 (Concourse)
Suntec Singapore - Level 3 Facilities (6-7 August 2009)



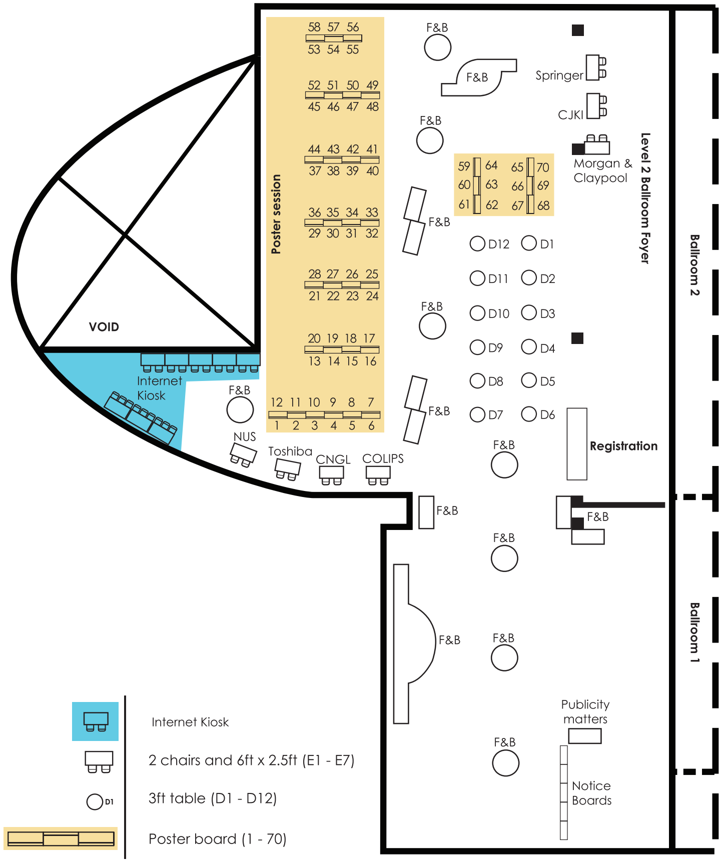
ACL-IJCNLP 2009 (Software Demos Session)

Suntec Singapore - Level 2 Ballroom Foyer
1730-1900hrs, 3 August 2009



ACL-IJCNLP 2009 (Poster Session-Short Papers & SRW Papers)

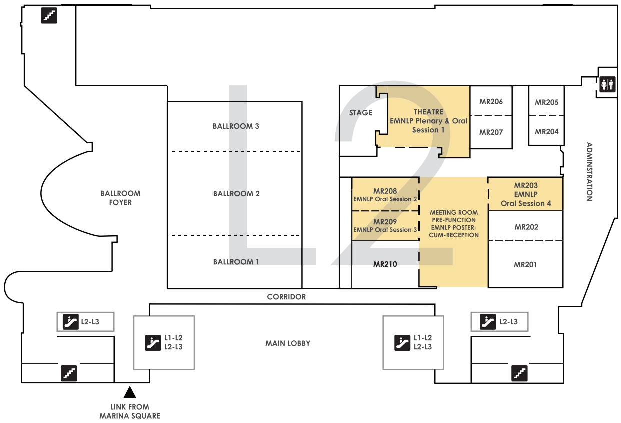
Suntec Singapore - Level 2 Ballroom Foyer
1230-1400hrs, 4 August 2009



EMNLP and Workshops

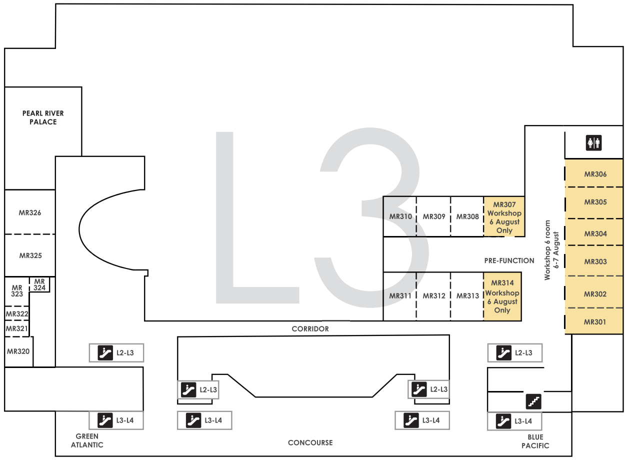
EMNLP Venue

Suntec Singapore - Level 2 Facilities (6-7 August 2009)



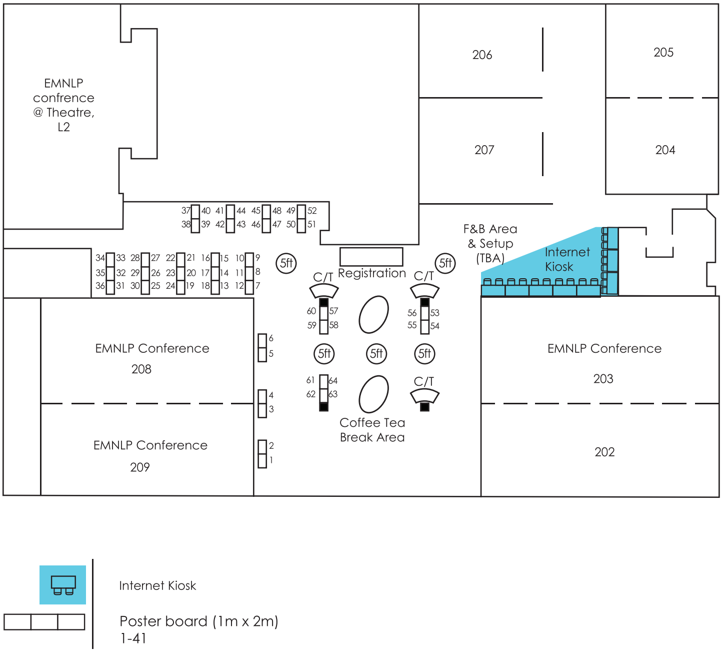
ACL-IJCNLP 2009 (Workshop)

Suntec Singapore - Level 3 Facilities (6 August only and 6-7 August)

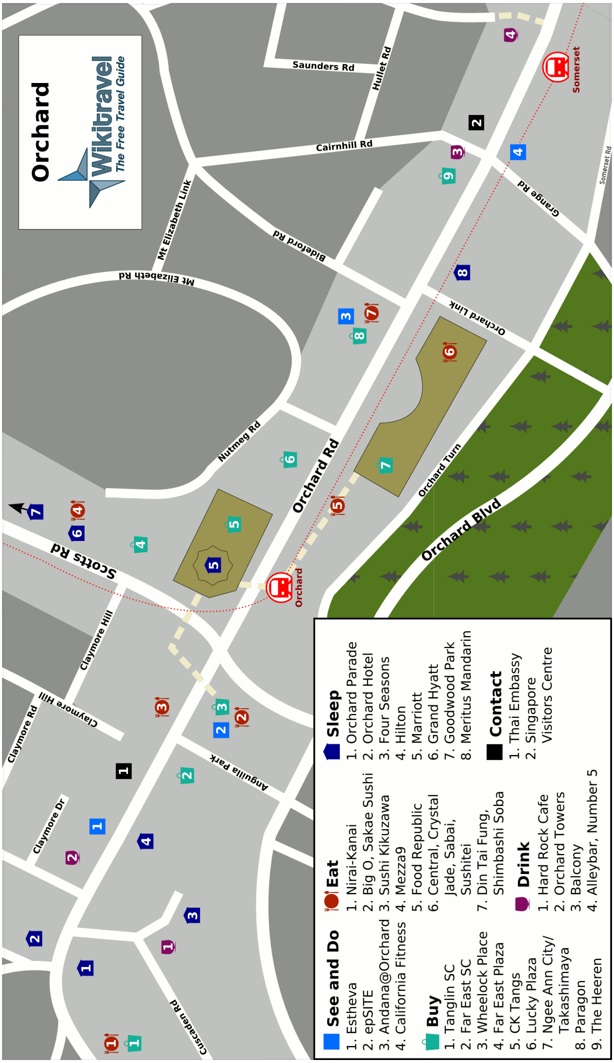


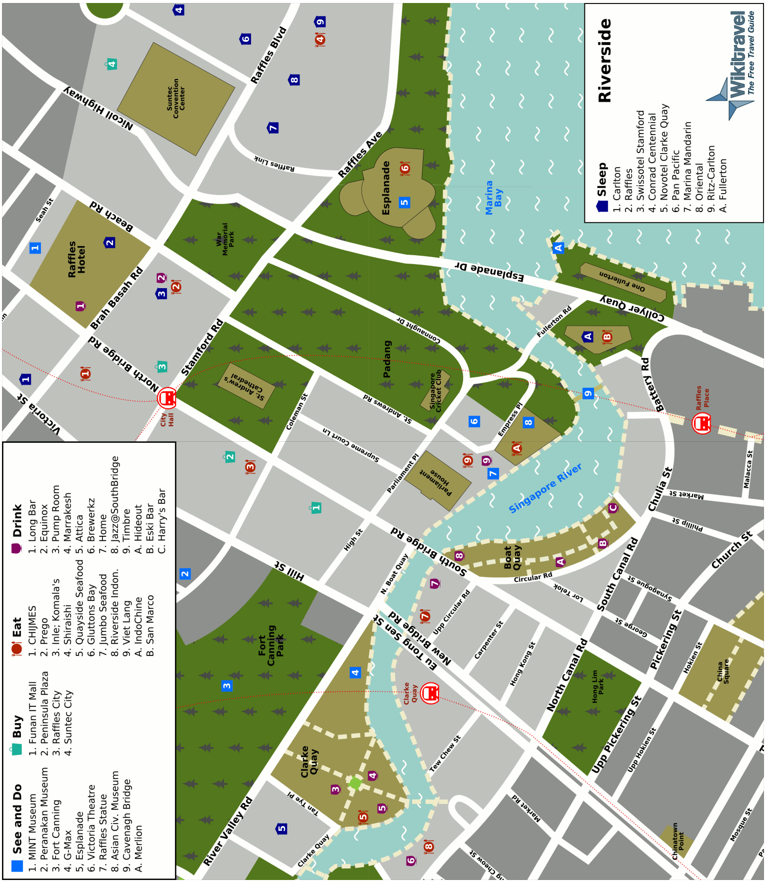
EMNLP 2009 (Poster-cum-Reception)

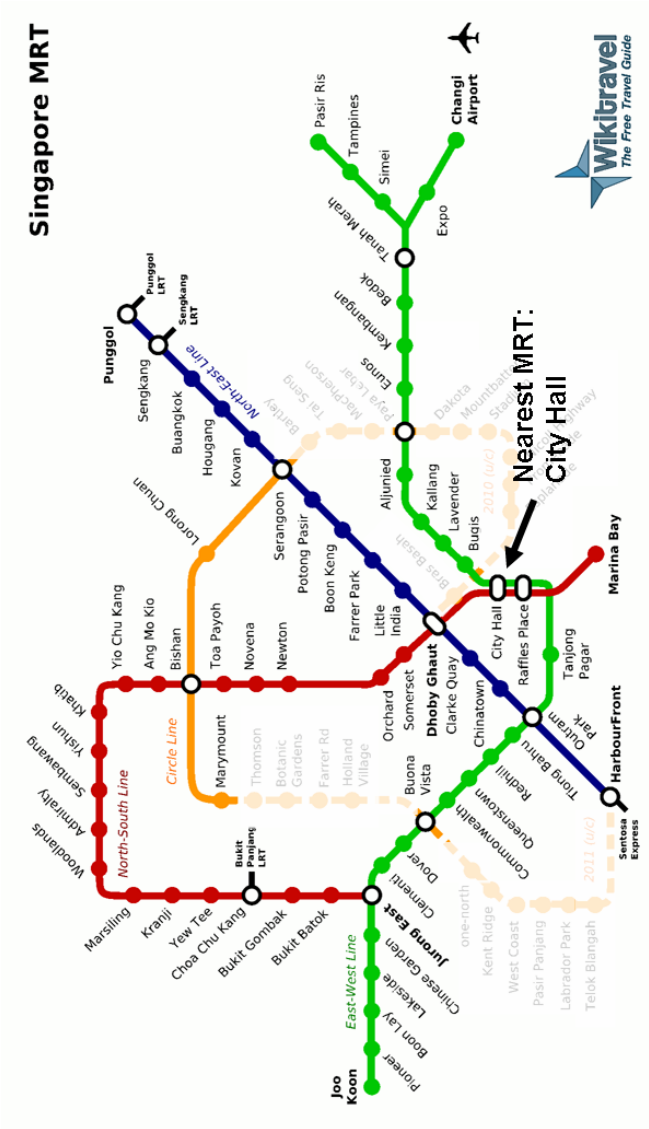
Suntec Singapore - Level 2 Meeting Room Foyer
1800-2000, 6 August 2009

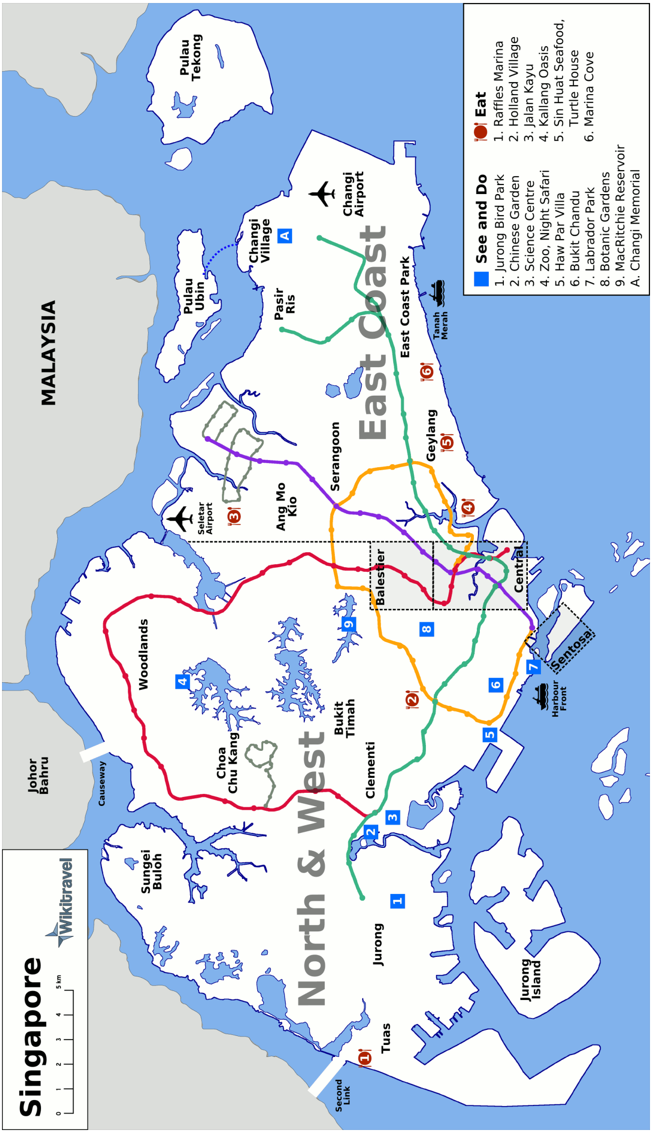


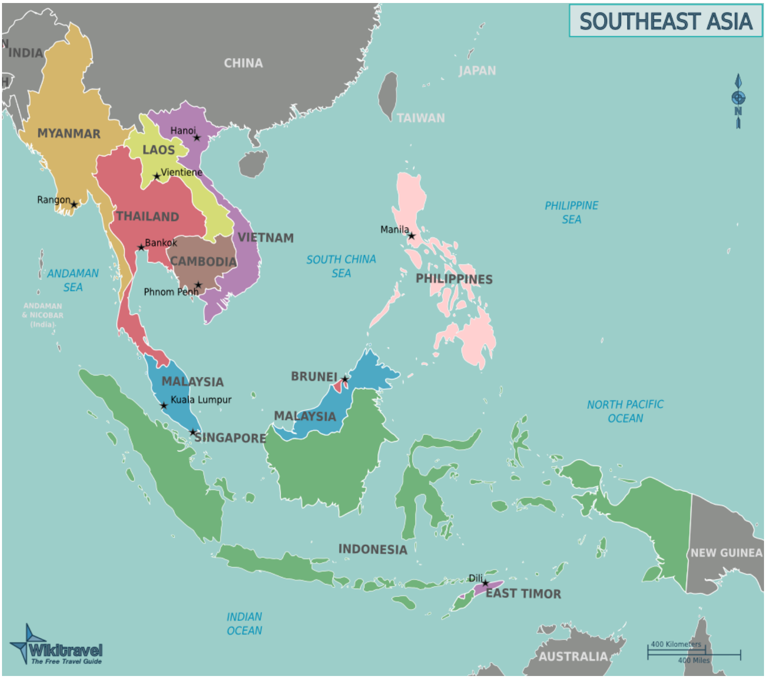
Singapore and Environs




Riverside
Sleep
1. Carlton
2. Raffles
3. Swissotel Stamford
4. Conrad Centennial
5. Novotel Clarke Quay
6. Pan Pacific
7. Marina Mandarin
8. Oriental
9. Ritz-Carlton
A. Fullerton
Wikitravel
The Free Travel Guide
See and Do
1. MINT Museum
2. Peranakan Museum
3. Fort Canning
4. G-Max
5. Esplanade
6. Victoria Theatre
7. Raffles Statue
8. Asian Civ. Museum
9. Cavenagh Bridge
A. Merlion
Buy
1. Funan IT Mall
2. Peninsula Plaza
3. Raffles City
4. Suntec City
Eat
1. CHIJMES
2. Prego
3. Inle; Komala's
4. Shiraishi
5. Quayside Seafood
6. Gluttons Bay
7. Jumbo Seafood
8. Riverside Indon.
9. Viet Lang
A. IndoChine
B. San Marco
Drink
1. Long Bar
2. Equinox
3. Pump Room
4. Marrakesh
5. Attica
6. Brewerkz
7. Home
8. Jazz@SouthBridge
9. Timbre
A. Hideout
B. Eski Bar
C. Harry's Bar
Nicoll Highway
Suntec Convention Center
Raffles Blvd
Raffles Link
Raffles Ave
Esplanade
Marina Bay
Seah St
Beach Rd
Raffles Hotel
Brah Basah Rd
War Memorial Park
Esplanade Dr
One Fullerton
Collyer Quay
North Bridge Rd
Stamford Rd
Victoria St
City Hall
St. Andrew's Cathedral
Coleman St
Connaught Dr
Padang
Singapore Cricket Club
Fullerton Rd
St. Andrews Rd
Supreme Court Ln
Empress Pl
Battery Rd
Raffles Place
Parliament Pl
Parliament House
Singapore River
Chulia St
Market St
Malacca St
Phillip St
Church St
High St
South Bridge Rd
N. Boat Quay
Boat Quay
Circular Rd
Hill St
Lor Telok
South Canal Rd
Synagogue St
George St
Upp Circular Rd
Eu Tong Sen St
New Bridge Rd
Carpenter St
Hong Kong St
Pickering St
Hokien St
China Square
Fort Canning Park
Clarke Quay
North Canal Rd
Hong Lim Park
Upp Pickering St
Yew Chew St
Upp Hokien St
River Valley Rd
Tan Tye Pl
Market Rd
Clarke Quay
Chinatown Point
Mosque St
Ng Cheow St


Singapore MRT
North-South Line
North-East Line
East-West Line
Circle Line
Jurong East
Bukit Batok
Bukit Gombak
Choa Chu Kang
Bukit Panjang LRT
Yew Tee
Kranji
Marsiling
Woodlands
Admiralty
Sembawang
Yishun
Khatib
Yio Chu Kang
Ang Mo Kio
Bishan
Toa Payoh
Novena
Newton
Orchard
Somerset
Dhoby Ghaut
City Hall
Raffles Place
Marina Bay
Punggol
Punggol LRT
Sengkang
Sengkang LRT
Buangkok
Hougang
Kovan
Serangoon
Potong Pasir
Boon Keng
Farrer Park
Little India
Clarke Quay
Chinatown
Outram Park
HarbourFront
Sentosa Express
Joo Koon
Pioneer
Boon Lay
Lakeside
Chinese Garden
Clementi
Dover
Buona Vista
Commonwealth
Queenstown
Redhill
Tiong Bahru
Tanjong Pagar
Bugis
Lavender
Kallang
Aljunied
Paya Lebar
Eunos
Kembangan
Bedok
Tanah Merah
Simei
Tampines
Pasir Ris
Expo
Changi Airport
Lorong Chuan
Marymount
Thomson
Botanic Gardens
Farrer Rd
Holland Village
one-north
Kent Ridge
West Coast
Pasir Panjang
Labrador Park
Telok Blangah
2011 (u/c)
Bartley
Tai Seng
MacPherson
Paya Lebar
Dakota
Mountbatten
Stadium
Nicoll Highway
Promenade
Esplanade
Bras Basah
2010 (u/c)
Nearest MRT:
City Hall
Wikitravel
The Free Travel Guide


Singapore
Wikitravel
0 1 2 3 4 5 km
MALAYSIA
Johor Bahru
Causeway
Second Link
Sungei Buloh
Woodlands
Choa Chu Kang
North & West
Bukit Timah
Clementi
Jurong
Tuas
Jurong Island
Seletar Airport
Ang Mo Kio
Serangoon
Pulau Ubin
Pulau Tekong
Changi Village
Pasir Ris
Changi Airport
East Coast
East Coast Park
Tanah Merah
Geylang
Balestier
Central
Sentosa
Harbour Front
1
2
3
4
5
6
7
8
9
A
See and Do
1. Jurong Bird Park
2. Chinese Garden
3. Science Centre
4. Zoo, Night Safari
5. Haw Par Villa
6. Bukit Chandu
7. Labrador Park
8. Botanic Gardens
9. MacRitchie Reservoir
A. Changi Memorial
Eat
1. Raffles Marina
2. Holland Village
3. Jalan Kayu
4. Kallang Oasis
5. Sin Huat Seafood, Turtle House
6. Marina Cove



Maps of Singapore and Environs courtesy Wikitravel.

14

Errata

- On page 7, the information for registering a WirelessSG account has recently changed. Registration confirmation and password are sent via SMS; local Singapore mobile phone numbers **are now required** at the moment (International phone numbers are currently not allowed). We advise you to purchase a prepaid SIM card upon arrival to Singapore. The prepaid SIM cards for all 3 mobile operators are sold in a number of shops including convenience stores like 7-Eleven. A valid passport is required to be shown in order to purchase a SIM card.

- On page 56, the student research workshop paper, *Optimizing Language Model Information Retrieval System with Expectation Maximization Algorithm*, is missing the second author, **Jyun-Wei Huang**.

- On pages 62 & 67, the *Future Conferences* and *Best Paper Awards* have been moved to the subsequent *Closing Session*.

- New, not in the handbook before. There is now a user account for all conference attendees to contribute your thoughts on the conference directly on the blog. Please visit

  http://www.colips.org/blog/acl-ijcnlp-2009/

  click the log in link on the right hand "Meta" menu, using "acl-ijcnlp" as the user name and "suntec" as the password.

  Once you've logged in, you will see the blog's dashboard, from which you can post your comments (via "posts") on the upper left. Please make sure to sign you own post and provide a link to your website if you wish.

  All posts will be reviewed on a (sub-) daily basis to ensure timely information gets posted. Commercial posts, advertisements, job openings related to ACL-IJCNLP are more than welcomed, but may be edited for content and delivery.