Morphological, Syntax and Semantic Knowledge in Statistical Machine Translation

http://sdrv.ms/11B3WZy

Marta R. Costa-jussà
Institute for Infocomm Research
Singapore

Chris Quirk
Microsoft Research
Seattle

About the speakers

Chris Quirk

- Senior Researcher at Microsoft Research
- Research areas include syntax-based translation, paraphrase
- BS (Carnegie Mellon 2000)

Marta R. Costa-jussà

- PhD (UPC, Barcelona, 2008)
- Research fellow at
  - LIMSI-CNRS (Paris),
  - Barcelona media,
  - Universidade de Sao Paulo,
  - Institute for Infocomm Research (Singapore)
- Main Research: Statistical Machine Translation

History of MT
MT approaches
    rule-based and corpus-based
SMT approaches
    phrase, syntax, hierarchical
SMT evaluation
Challenges and applications

OUTLINE

HISTORY OF MT

Do we need to motivate MT?

MT is commercially and academically interesting

Commercially	Academically
Military and intelligence purposes	Study the basic mechanisms of language and mind
MT is popular on the web	MT requires from other NLP technologies: parsing, generation, word sense disambiguation, named entity recognition, transliteration, pronoun resolution, real-world knowledge…
Transmission of technical, agricultural and medical information to the developing countries	Exploit the power of the computer

Georgetown-IBM Experiment

New York, January 7, 1954: Russian was translated into English by an electronic "brain" today for the first time. […]

MT, one of the first applications evisioned for computers

- 1947-1954 Information Theory Foundations
- 1954-1966 Different levels of representation explored
- 1966-1980 ALPAC Report, after that, research continued in Europe and Canada
- 1980s Variety of systems: rule-based interlingua and data-driven. First commercial systems
- 1990s Statistical Machine Translation
- 2000s MT software

What is going on now in MT?

- MT is consistently improving with resources and computational power
- Reasonable quality when resources available
- Moving to Hybrid Architectures with Statistical and Linguistic knowledge to cover what data does not cover

Rule-based and Corpus-based

MT APPROACHES

Approaches to Machine Translation

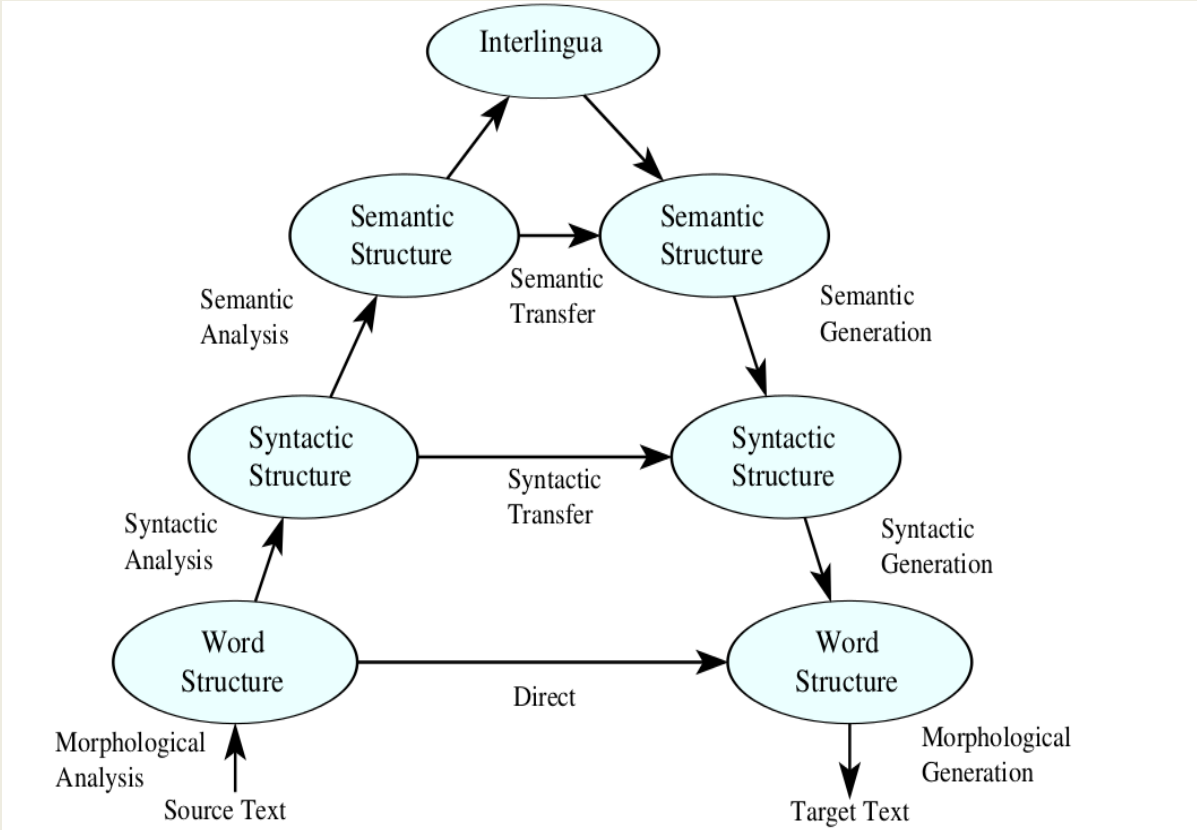
Sources of information

Rule-based: human written specific rules

Corpus-based: use data to learn

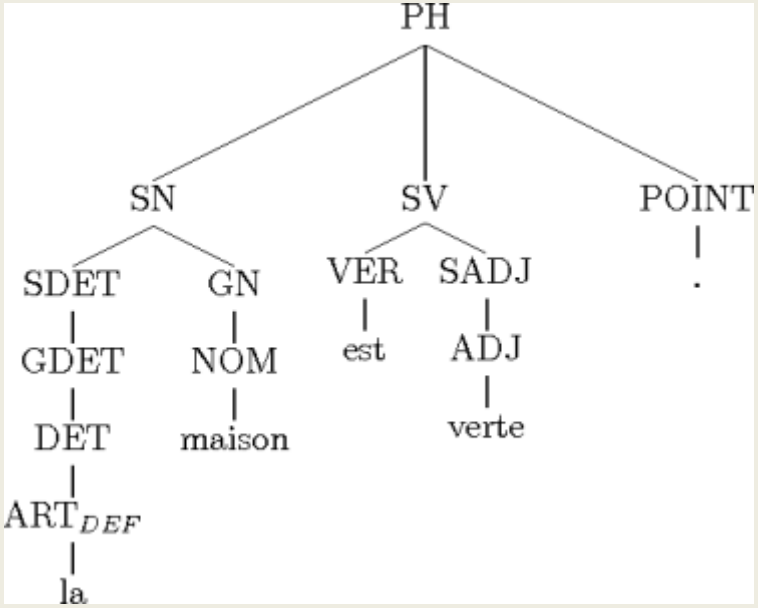
Level of representation

Transfer-based MT has several depths of intermediary representation



Microsoft Research

Rule-based Machine Translation



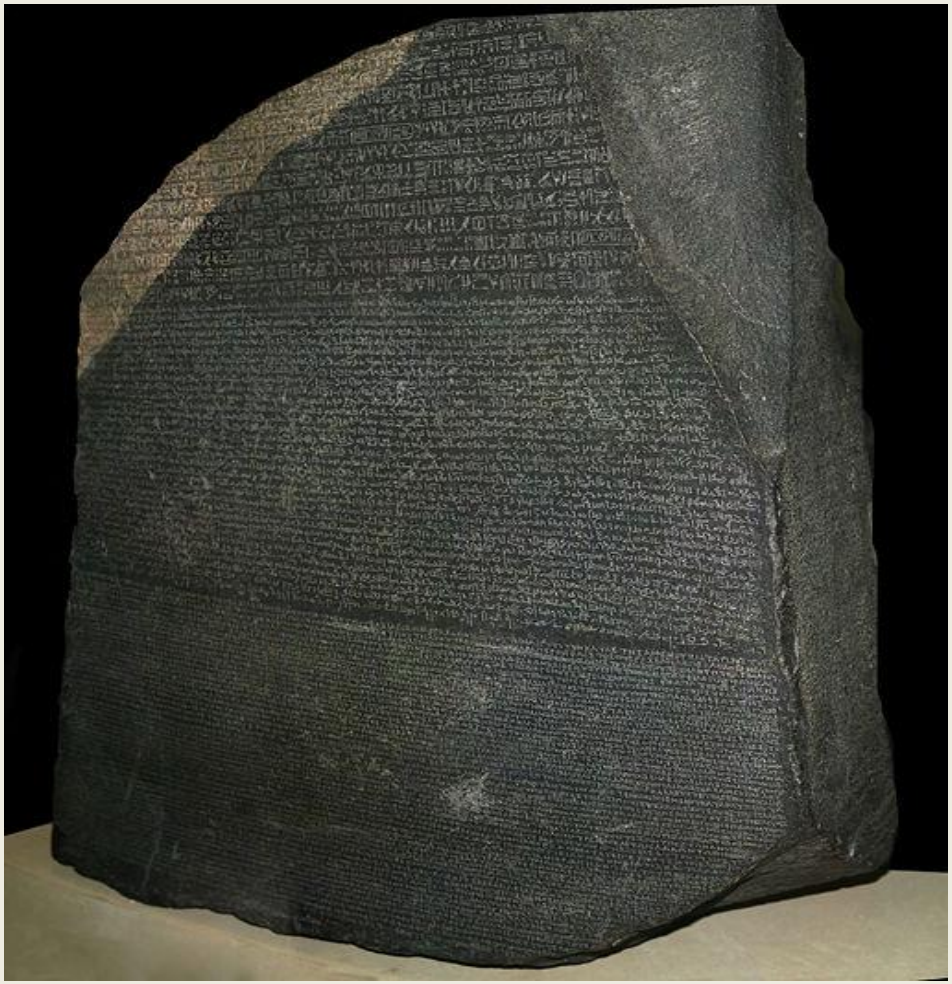
- Resources:
  - Morphological dictionaries
  - Source parser
  - Translation lexicon
  - Transfer rules

Corpus-based MT are trained on parallel corpora

Collections of parallel texts at sentence level

English	Russian
This course is a thorough introduction to machine translation technology	Этот курс представляет собой интенсивное введение в технологию машинного перевода
We will describe all aspects of building a statistical machine translation system, from both formal and practical perspectives	Мы рассмотрим все аспекты построения системы статистического машинного перевода с теоретической и практической точки зрения

An early parallel text



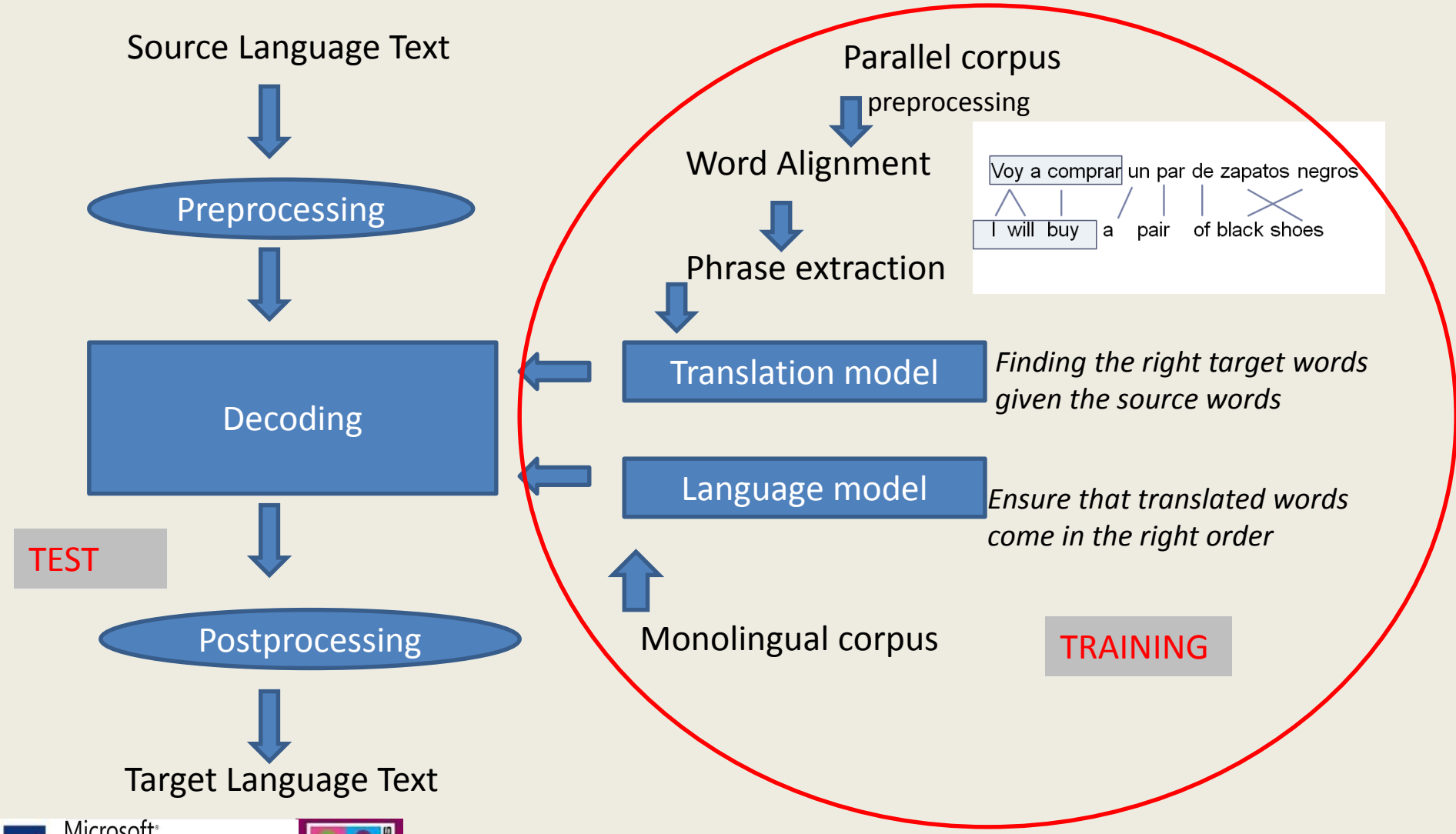
Advantages of SMT

- Data driven
- Language independent
- No need for staff of linguists of language experts
- Can prototype a new system quickly and at a very low cost
- High flexibility of matching heuristics
- High coverage

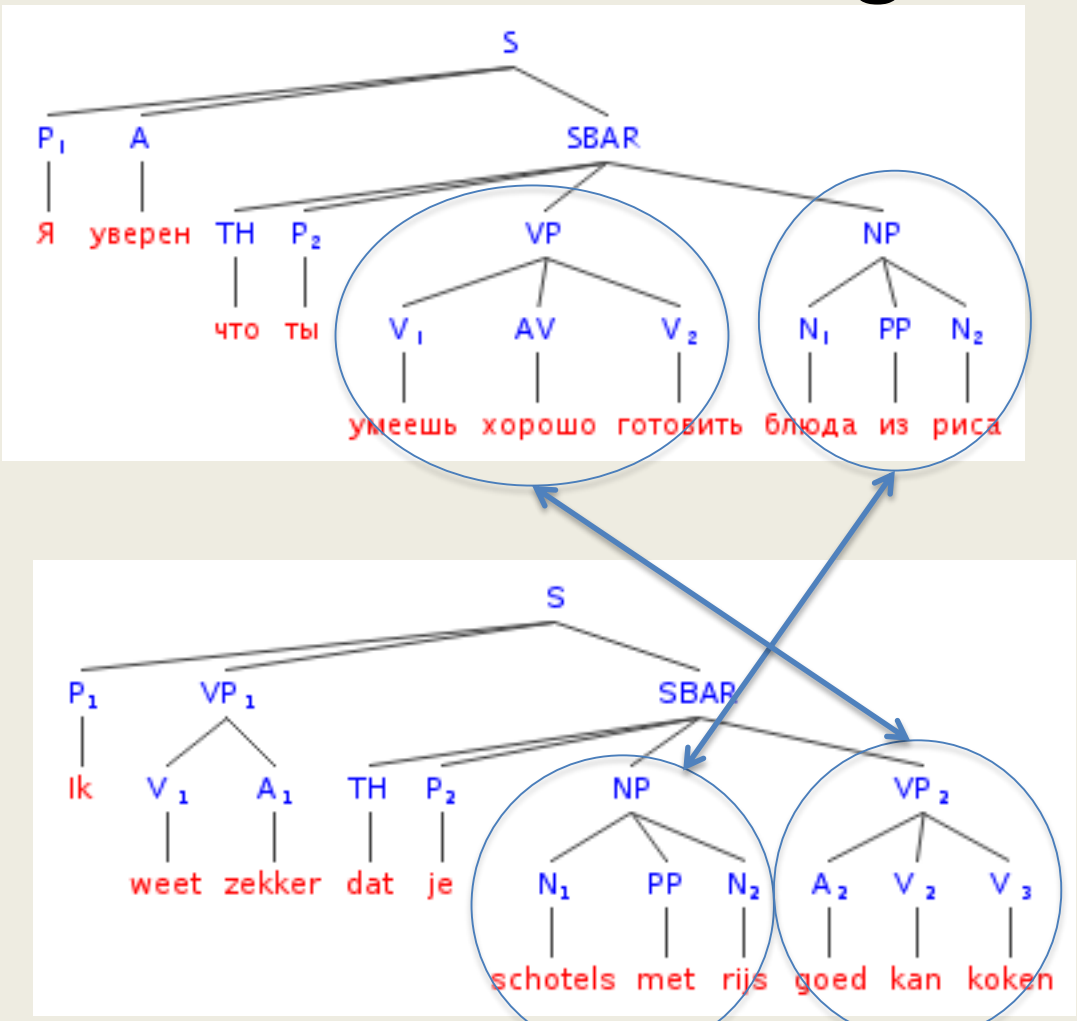
Phrase, syntax and hierarchical

SMT APPROACHES

A picture is worth a million equations



Syntax Augmented introduce syntax trees in decoding



Hierarchical-based introduce hierarchical rules in decoding

- Hierarchical rules allow for hierarchical phrases that can contain other phrases

[Я] [уверен] [что] [ты] [хорошо умеешь готовить] [блюда с рисом]

[Ik] [weet zeker] [dat] [je] [schotels met rijst] [goed kan koken]

[ты][X][блюда с рисом] - > [je][schotels met rijst][X]

SMT EVALUATION

SMT Evaluation

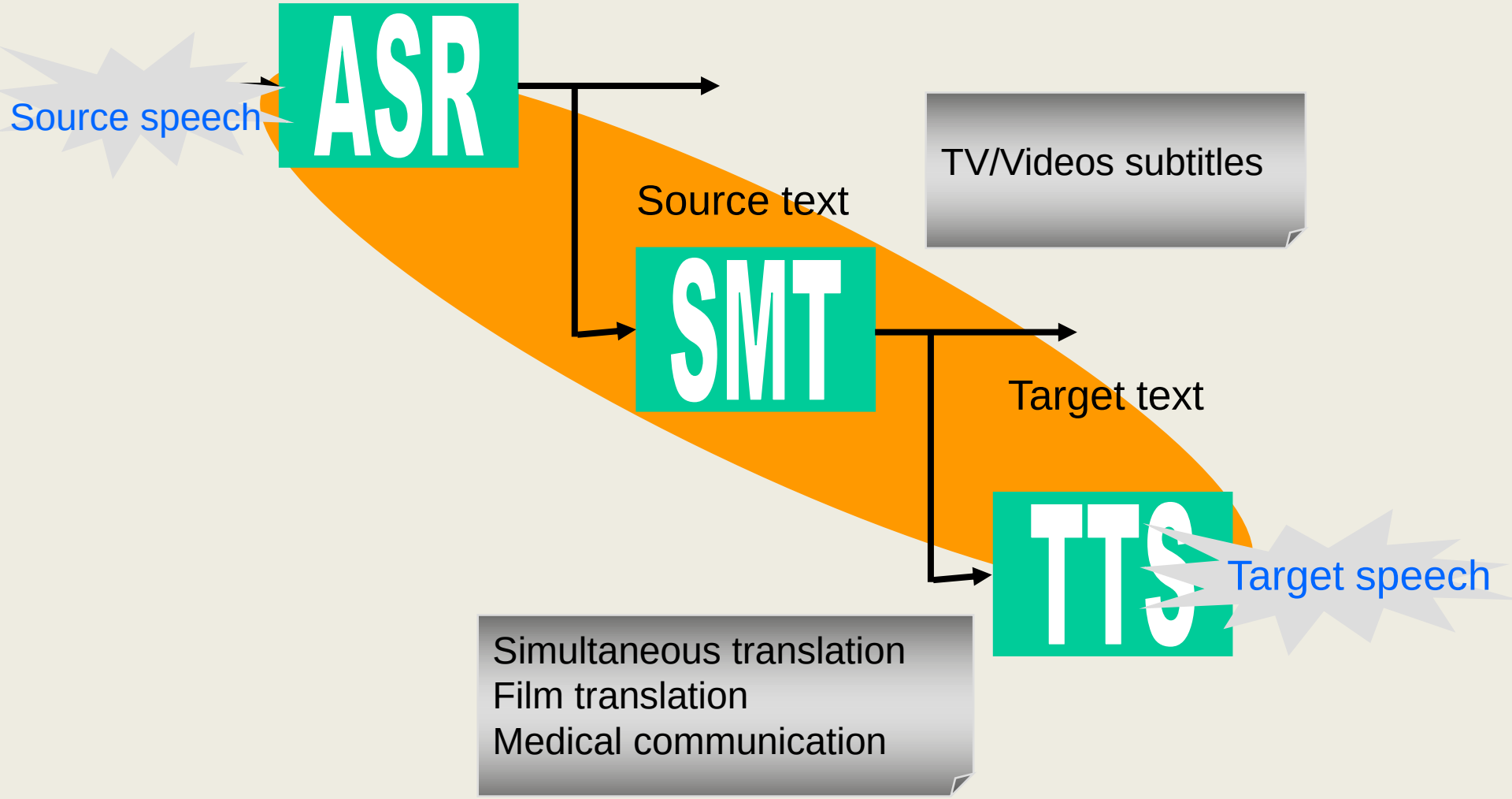
- Automatic evaluation allows to optimize the systems.
  - BLEU is the standard measure taken by the scientific comunity. It evaluates sequences of ngrams
- Human evaluation allows for a fair comparison across systems
  - FLUENCY, ACCURACY are the two standard measures

CHALLENGES AND APPLICATIONS

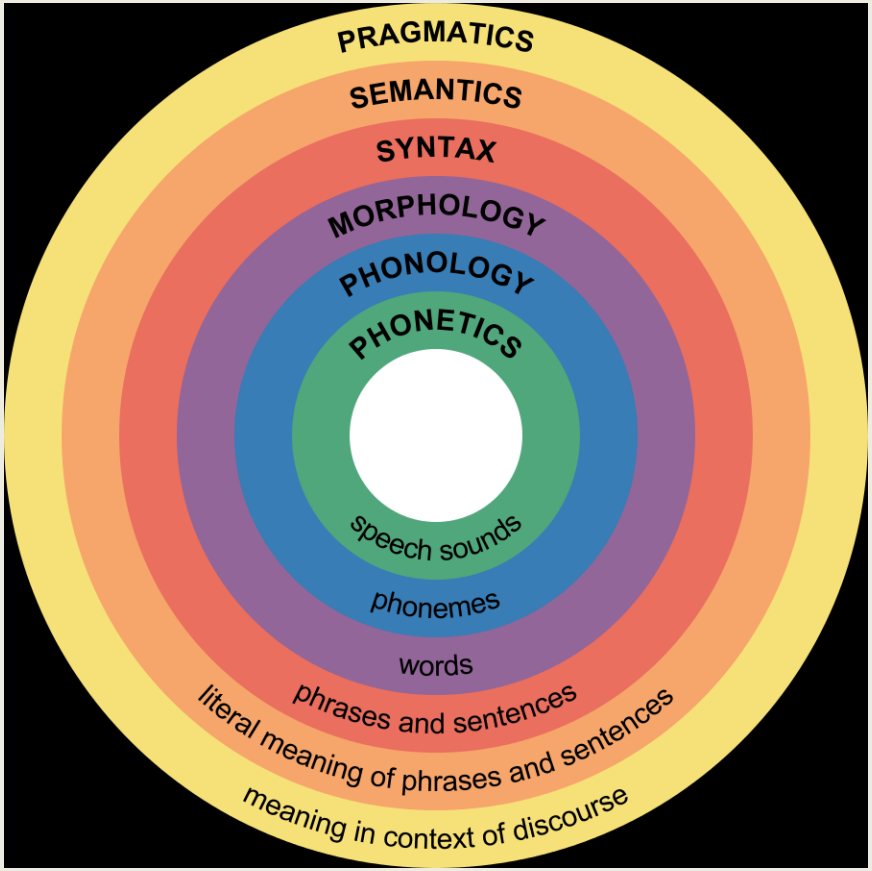
MT performance quite acceptable in …

- CLIR allows to search for hotels in Paris on French-language pages or bars in Moscow from Russian sources.
- Computer-aided human translation
- Communication via email, chat
- Translation on hand-held devices

Needs to further improve …



SMT challenges
Language linguistics



SMT challenges are found at all linguistic levels

- Morphology: word forms

  You/We/They are ---- Vôce/Nos/Vôces é/somos/são

- Syntax: word order

  SVO --- SOV

- Semantics: word sense, idioms

  banco --- bank, sit???

What are recent methods of including **morphological**, **syntax** and **semantic** knowledge in **SMT**?

Tutorial overview

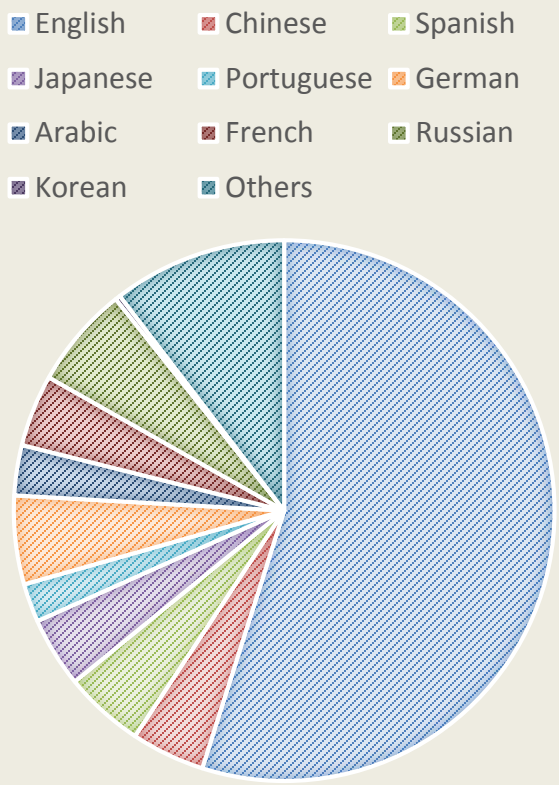
- Morphology in SMT
- Syntax in SMT
- Semantics in SMT


A*STAR
Microsoft
Research
MARIE CURIE
ACTIONS

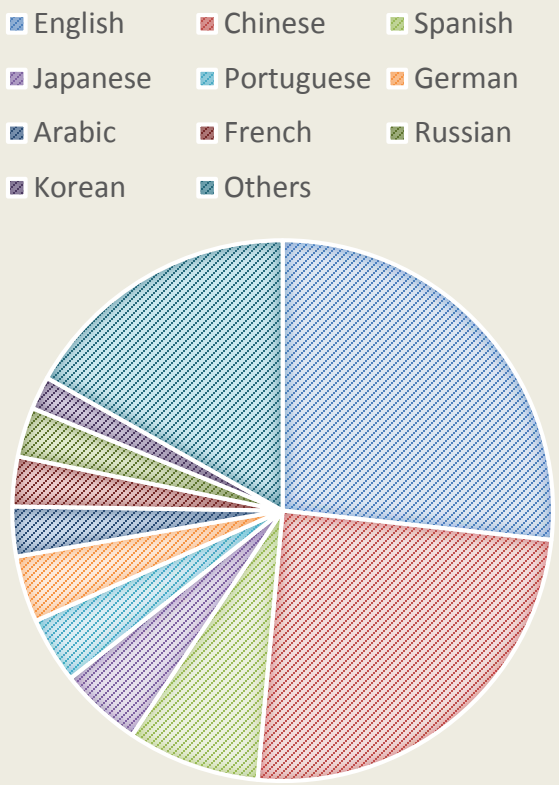
Usage statistics

http://en.wikipedia.org/wiki/Languages_used_on_the_Internet

Internet ***pages*** by language



Internet ***users*** by language



Note the huge mismatch!

Literal mindedness

- English to Russian

The spirit is willing but the flesh is weak

- Russian to English

The vodka is good but the meat is rotten

Morphology knowledge in Statistical Machine Translation

Marta R. Costa-jussà

Chris Quirk

TUTORIAL NAACL 2013

Morphology concepts

Morphology-related challenges in SMT

Segmentation approaches

Generation approaches

Enriching approaches

Conclusions

OUTLINE

Word structure + formation

viviendo vivir

terbiye terbiyesiz

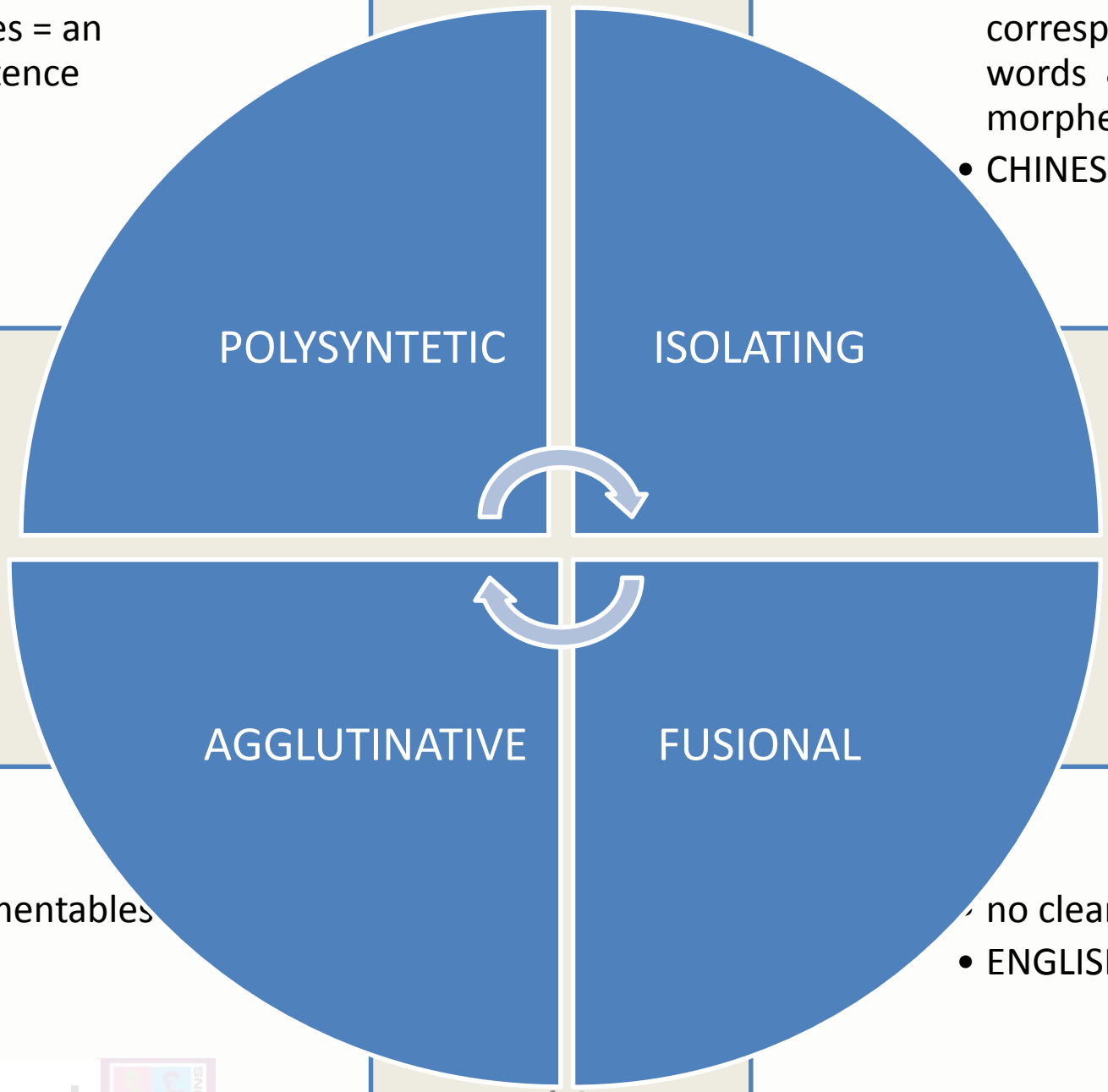
houses house

Common morphological operations

- AFFIXATION: *nation + al*
- COMPOUNDING: *sun+glasses*
- REDUPLICATION: *bye-bye*
- INTERNAL CHANGE: *rang [instead of ringed]*
- SUPPLETION: *went [past of go*]
- BLENDING*: motel [motor+hotel]*

- a word of several morphemes = an entire sentence
- INUIT

- one-to-one correspondance words & morphemes.
- CHINESE

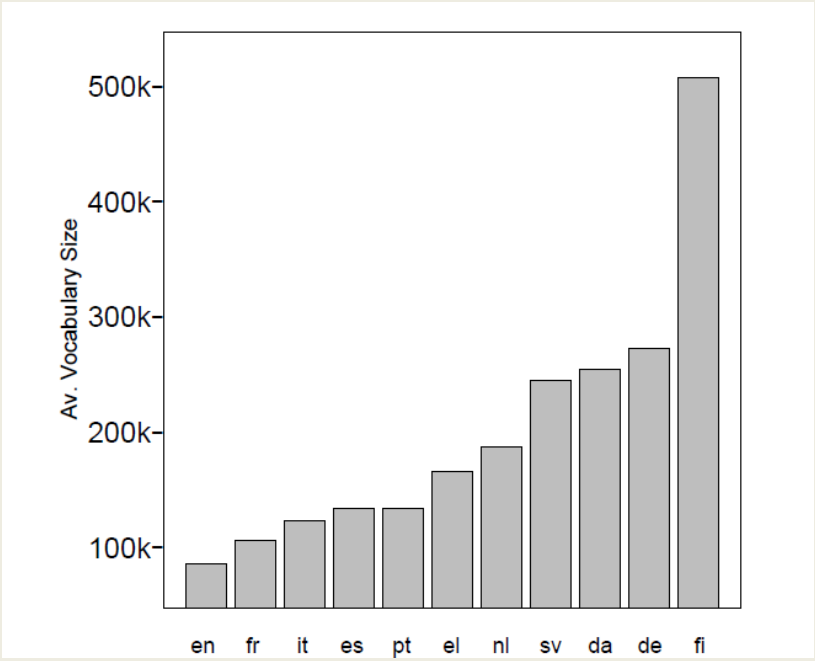


- easily segmentables
- TURKISH

- no clear boundaries
- ENGLISH

MORPHOLOGY-RELATED CHALLENGES IN SMT

Sparsity



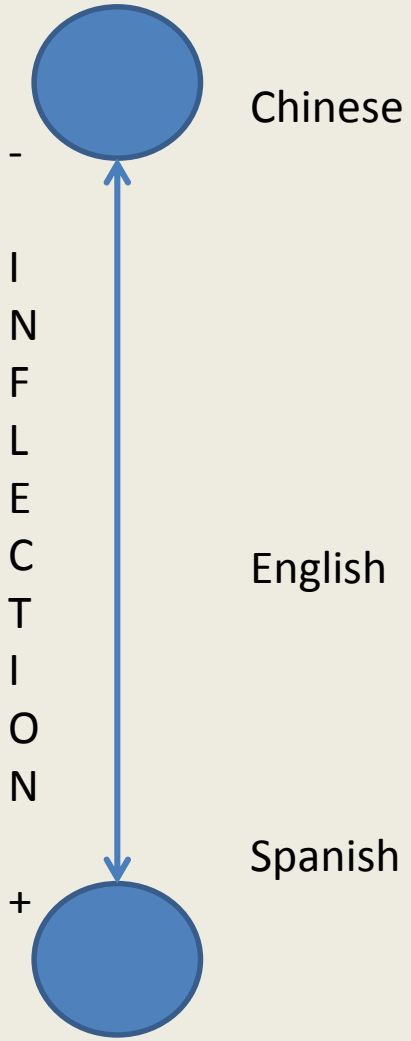
tietä+isi+mme

know+would+we

Creutz et al. 2005

Morphology mapping

- Challenges
  - Lack of information to generate the correct word form



isolating ↔ fusional/agglutinative

- Isolating language

  是

- High-inflected language
  - Yo soy -Nosotros somos
  - Tu eres -Vosotros sois
  - Él es -Ellos son

Chinese ↔ Spanish

Long distance agreement error

REF: Maria is buying her first house

MT: Maria is buying his first house

MORPHOLOGY APPROACHES TO ADDRESS MORPHOLOGY IN SMT

high → low inflected

- Preprocessing techniques
  - Segmentation approaches

“easy” task

from big to small space

low → high inflected

- Postprocessing techniques
  - Generation
  - Enriching models

difficult task

from small to big space

Make vocabularies more similar

SEGMENTATION

Language-dependent segmentation

- English into Spanish/Catalan task:
  - Treatment of verbs: identify (by means of POS) pronoun+verb sequence and splice these two words into one,
    - » you go --- PRP VBP --- you_go
- Spanish/Catalan into English task:
  - split contractions (e.g. del = de + el, al =a +el)

Ueffing et al. 2003

Language-dependent segmentation

- Arabic-to-English task.

TOK	
ST	Splitting off punctuation and numbers
D1	Declitization (w+, f+)
D2	Declitization (D1+ l+, k+, b+, s+)
D3	Declitization (D1,D2, Al+)
MR	Stem + affixival morphemes
EN	English-like

Habash et al, 2006

Language-dependent segmentation

- Arabic-to-English task.

Input	wsynhY	Alr}ys	jwlth	bzyArp	AlY	trkyA.	
Gloss	and will fi nish	the president	tour his	with visit	to	Turkey	.
English	The president will fi nish his tour with a visit to Turkey.						
ST	wsynhY	Alr}ys	jwlth	bzyArp	AlY	trkyA	.
D1	w+ synhy	Alr}ys	jwlth	bzyArp	<lY	trkyA	.
D2	w+ s+ ynhy	Alr}ys	jwlth	b+ zyArp	<lY	trkyA	.
D3	w+ s+ ynhy	Al+ r}ys	jwlp +P_{3MS}	b+ zyArp	<lY	trkyA	.
MR	w+ s+ y+ nhy	Al+ r}ys	jwl +p +h	b+ zyAr +p	<lY	trkyA	.
EN	w+ s+ >nhY$_{VBP}$ +S_{3MS}	Al+ r}ys$_{NN}$	jwlp$_{NN}$ +P_{3MS}	b+ zyArp$_{NN}$	<lY$_{IN}$	trkyA$_{NNP}$	.

  - Small data set: English-like tokenization
  - Large data set: splitting only some clítics

Habash et al, 2006

Language-independent segmentation

- Morfessor is a method for finding morpheme-like units of a language in an unsupervised manner.
  - Minimum Description Length

Example of segmentation:

affectionate affect+ion+ate

Creutz et al, 2005
Virpioja et al., 2007

Language-independent segmentation

- Categories-ML: probabilistic model, unsupervised manner, with categories: prefix, stem and suffix
- Categories-MAP introduces a hierarchical lexicon structure with categories: prefix, stem and suffix + NON category used for internal representation
- Specially effective for agglutinative languages

Cover cases when correct translations are not in the phrase table

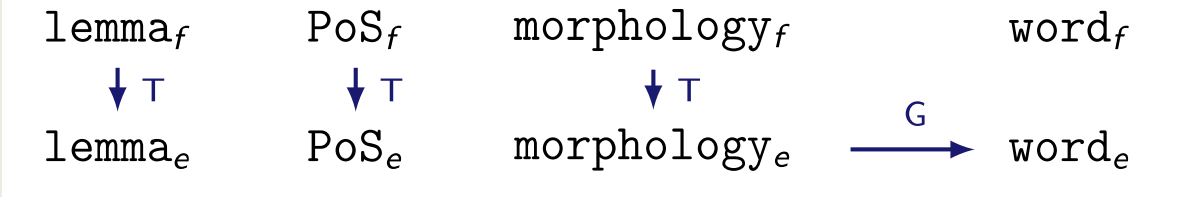
GENERATION

Factored translation models

- Factored translation models are an extension to phrase-based models where every word is substituted by a vector of factors.

(word) ==⇒ (word, lemma, PoS, morphology, ...)

- The translation is now a combination of pure translation (T) and generation (G) steps:



Koehn et al., 2007

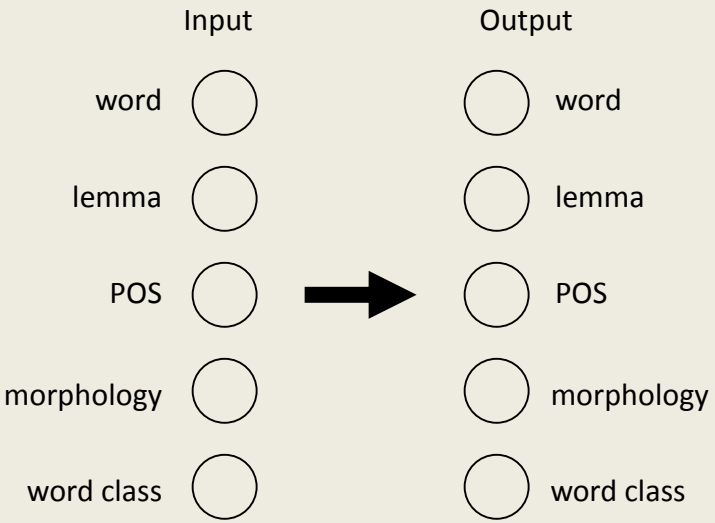
Factored translation models

What differs in factored translation models
(as compared to standard phrase-based models)

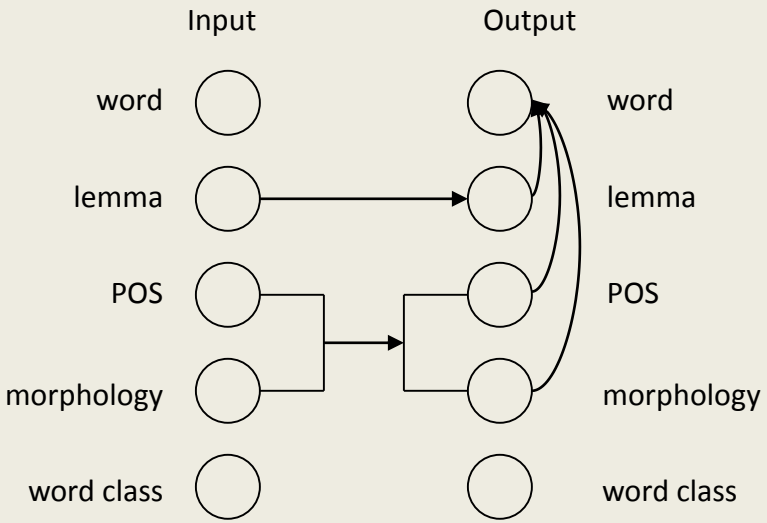
- The parallel corpus must be annotated beforehand.
- Extra language models for every factor can also be used.
- Translation steps are accomplished in a similar way.
- Generation steps imply a training only on the target side of the corpus.
- Models corresponding to the different factors and components are combined in a log-linear fashion.

Factored translation models

Factored Representation



Factored Model: transfer and generation



Automatic Post Edition approaches

Simplified translation in terms of morphology and automatic post edition of inflection

1. Different types of simplification: PoS generalization, Stems
2. Automatic post edition by means of SVM or Maximum Entropy

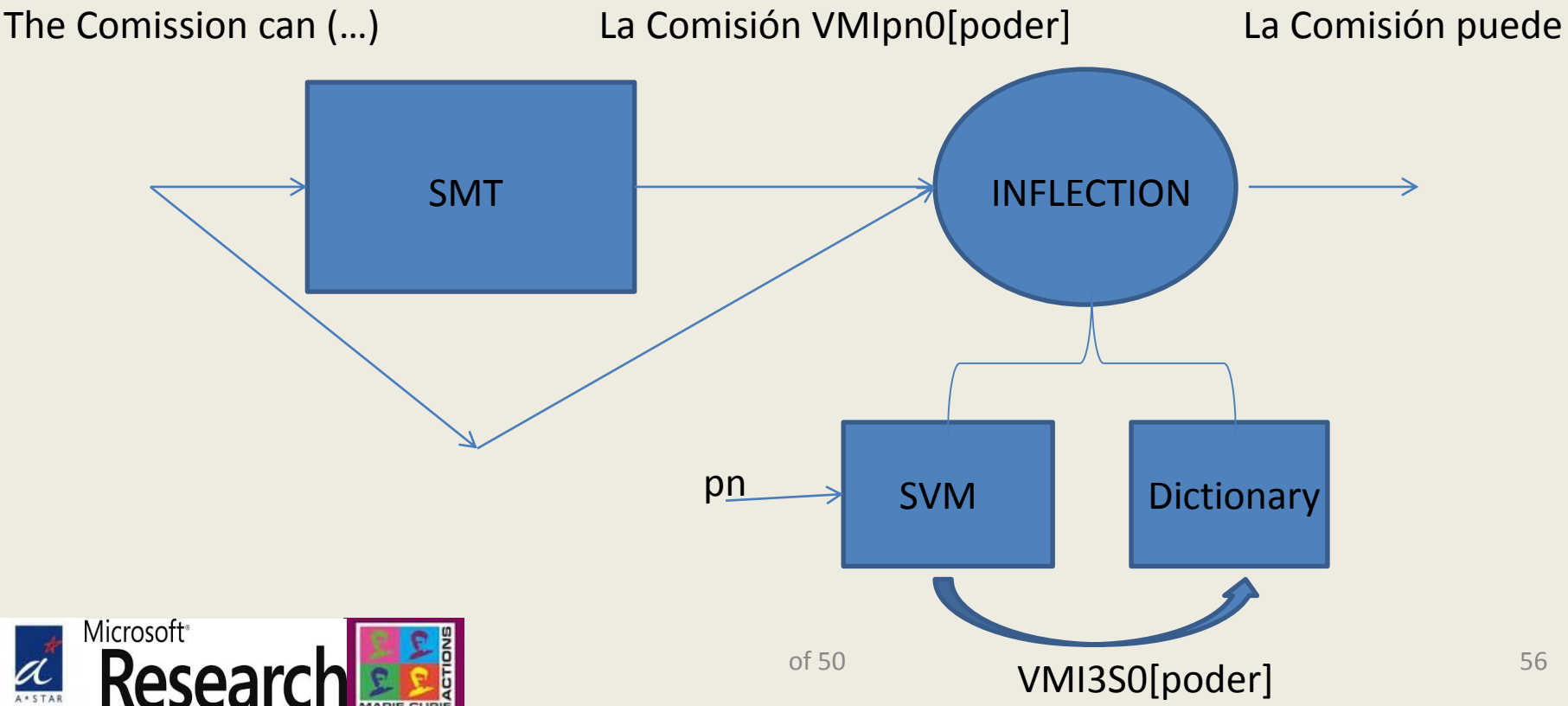
PoS verb morphology simplification

Type	Text
Plain target	La Comisión puede llegar a paralizar el programa
Lemma + PoS	La Comisión VMIP3S0[poder] llegar a paralizar el programa
Lemma+PoS Generalized	La Comisión VMIpn0[poder] llegar a paralizar el programa

1st APE approach

Formiga et al. 2012

SMT + APE (verb inflection)



SVM and DDAG multi-class

Person and Number — Number and Gender
Developed under the **framework of a Multiclass Classification System.**

Two parallel prediction tasks: (Person and Number — Number and Gender):

1 Non-participle: 6 output classes (1st / 2nd / 3rd either Sing./Plural)
2 Participle/adjectives from verb : 4 output classes (Male/Female either Sing./Plural)

Source/Target Features
Context words and PoS
Verb phrase with POS
Presence of passive voice
Named entities
Reflexive pronoun
Pronouns before the source and target verbs with POS

Translation to stems (for all categories)

Type	Text
Plain target	La Comisión puede llegar a paralizar el programa
Stem	la comisión pued lleg a paraliz el program

- Produce set of all possible stems for a word w

2nd APE approach

Toutanova et al., 2008

Inflection over all word forms

- In this case, inflection returns the set of surface word forms for the stems according to the lexicon
  - On average 14 word forms per stem in Russian

Morphological Analysis

- The morphological analysis returns the set of possible morphological analyses for a word.
  - Features for Russian: POS, person, Number, Gender, Tense, Mood and Case

Models for Inflection Prediction

- The task*: given a source sentence, a sequence of stems in the target language and additional morpho-syntactic annotations, select an inflection from its inflection set for every stem*
- Maximum Entropy
- The model implemented is of second order

$$p(\bar{y}|\bar{x}) = \prod_{t=1}^{n} p(y_t|y_{t-1}, y_{t-2}, x_t), y_t \epsilon I_t$$

Large phrase tables already contain many word forms

ENRICHING MODELS

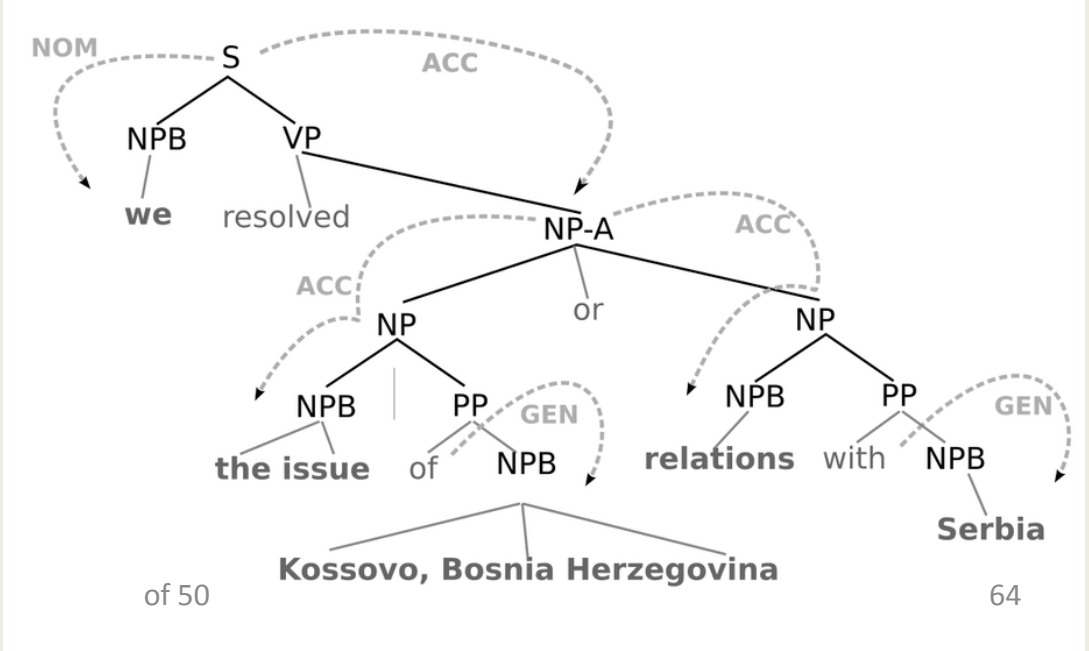
Enriching source text

- EN: The president, after reading the press review and the announcements, left his office
- GR-1: The president[nominative], after reading[3S] the press review[Accusative,S] and the announcements[***Accusative***,p], left[3S] his office[Accusative,S]
- GR-2: The president[nominative], after reading[3S] the press review[Accusative,S] and the announcements[***Nominative***,p], left[3S] his office[Accusative,S]

Avramadis et al. 2008

Noun cases agreement

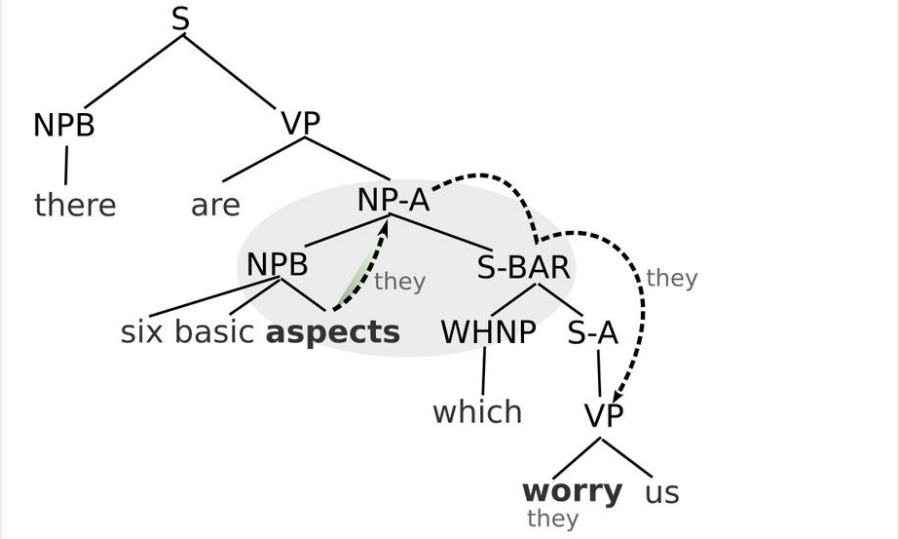
- Use of syntactic role
  - Nominative case, subject
  - Accusative case, direct object
  - Dative, indirect object



Verb person conjugation

- Identify the person of a verb (subject selection)
  - Directly inferred by a personal pronoun
  - Pronouns in a different case (them, myself) where converted into nominative
  - Subject is a single noun means 3rd person, POS tag is used to identify S or P
  - Gender does not affect

- Appying person tags



Morphological Word Classes

- In addition to the word forms
  - POS target language model
  - Statistical classes learned with clustering
  - Class-based agreeement model

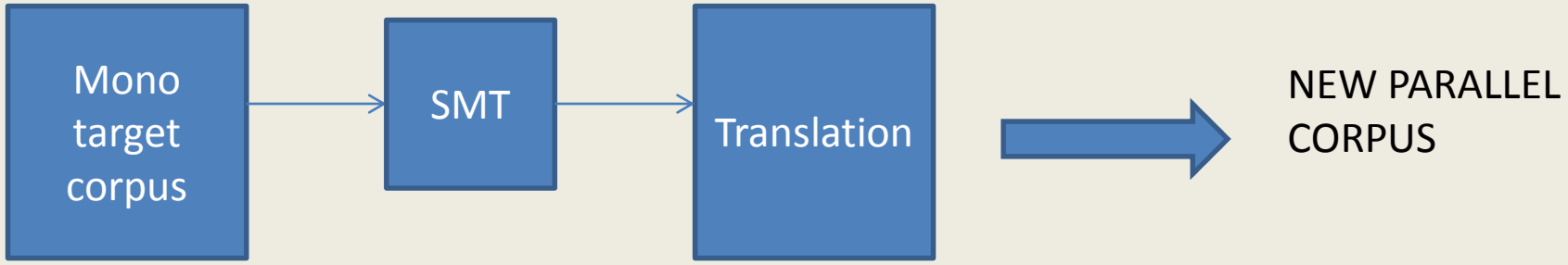
noun+fem+sg verb+fem+sg+3

N(+F) V(+F)

Green et al, 2012

Reverse self-training with back-off

- **Translate** target monolingual corpus and add as training parallel corpus.



- Translation of the monolingual corpus is done by using the target either the word form or a **simplified version** of it (i.e. lemmas).

Bojar et al, 2011

Learning Unseen Forms

Small Parallel Data

Source	Target	Target Lemma
A cat chased	**kočka honila...**	**kočka honit...**
I saw a cat	**kočku vidět**	**být kočka**
I read about a dog	četl jsem o psovi	číst být o pes

Large Monolingual Data:

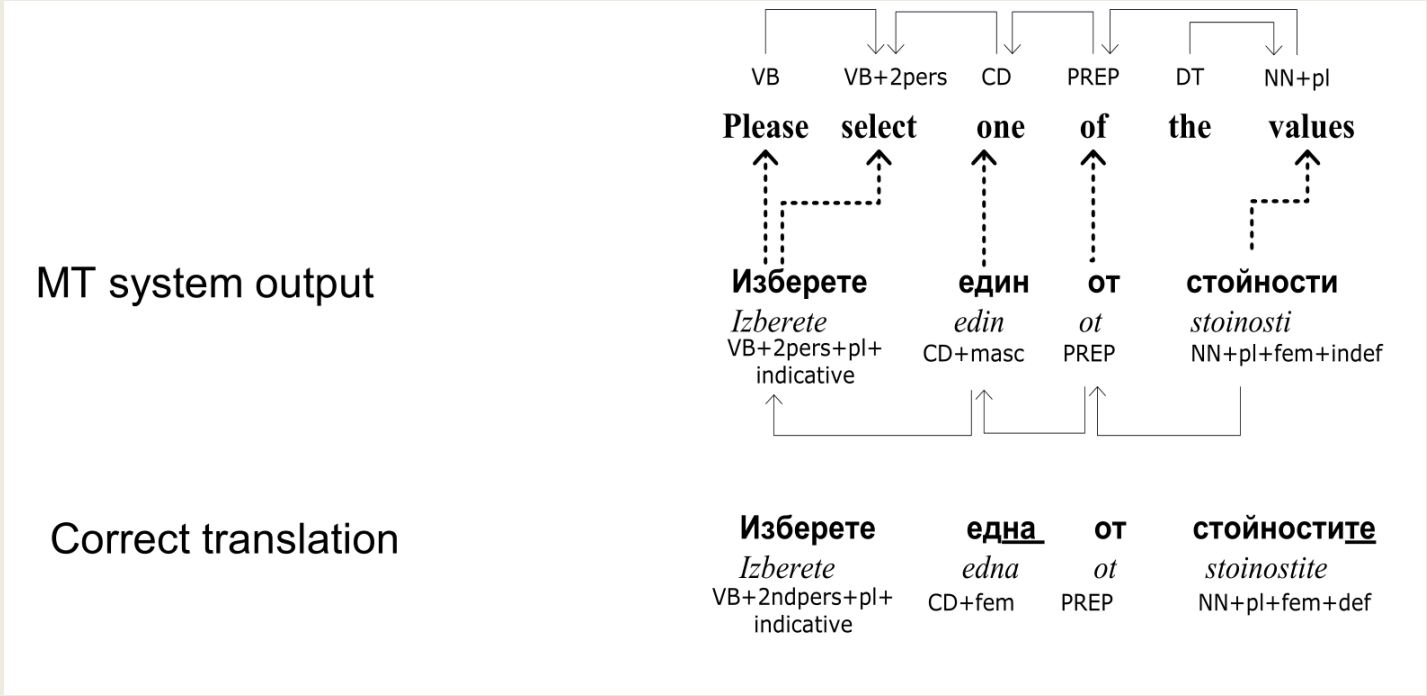
Source	Target	Target Lemma
?	**četl jsem o kočce**	**číst být o kočka**

I read about a cat ¬ Use reverse translation backed-off by lemmas

- Learned a new phrase (**o kočce**) including a form never seen in parallel data (**kočce**).

Discriminative selection models

- Better lexical selection, especially for morphologically complex languages



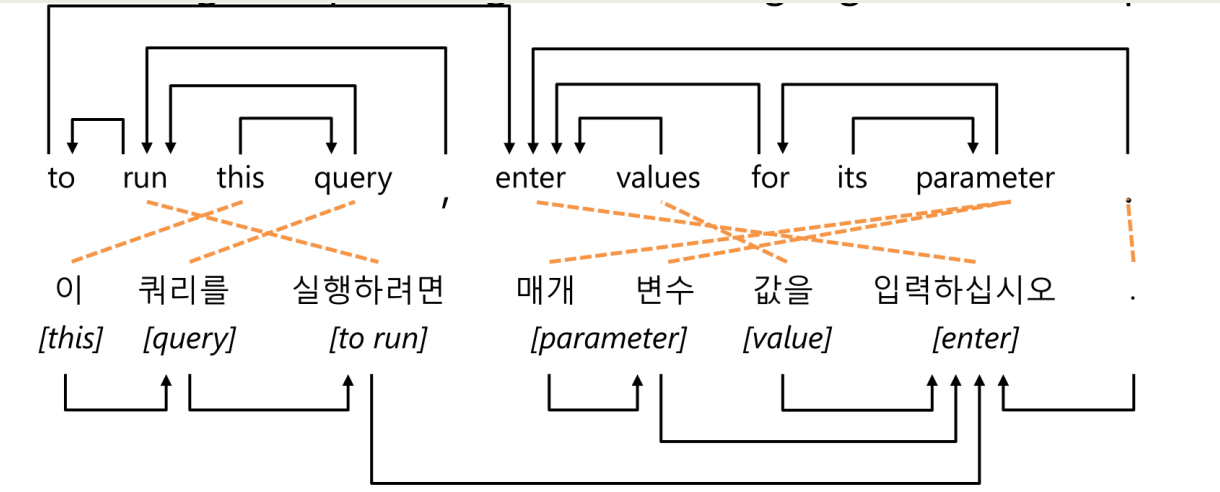
Jeong, Toutanova, Suzuki, and Quirk 2010

Approach

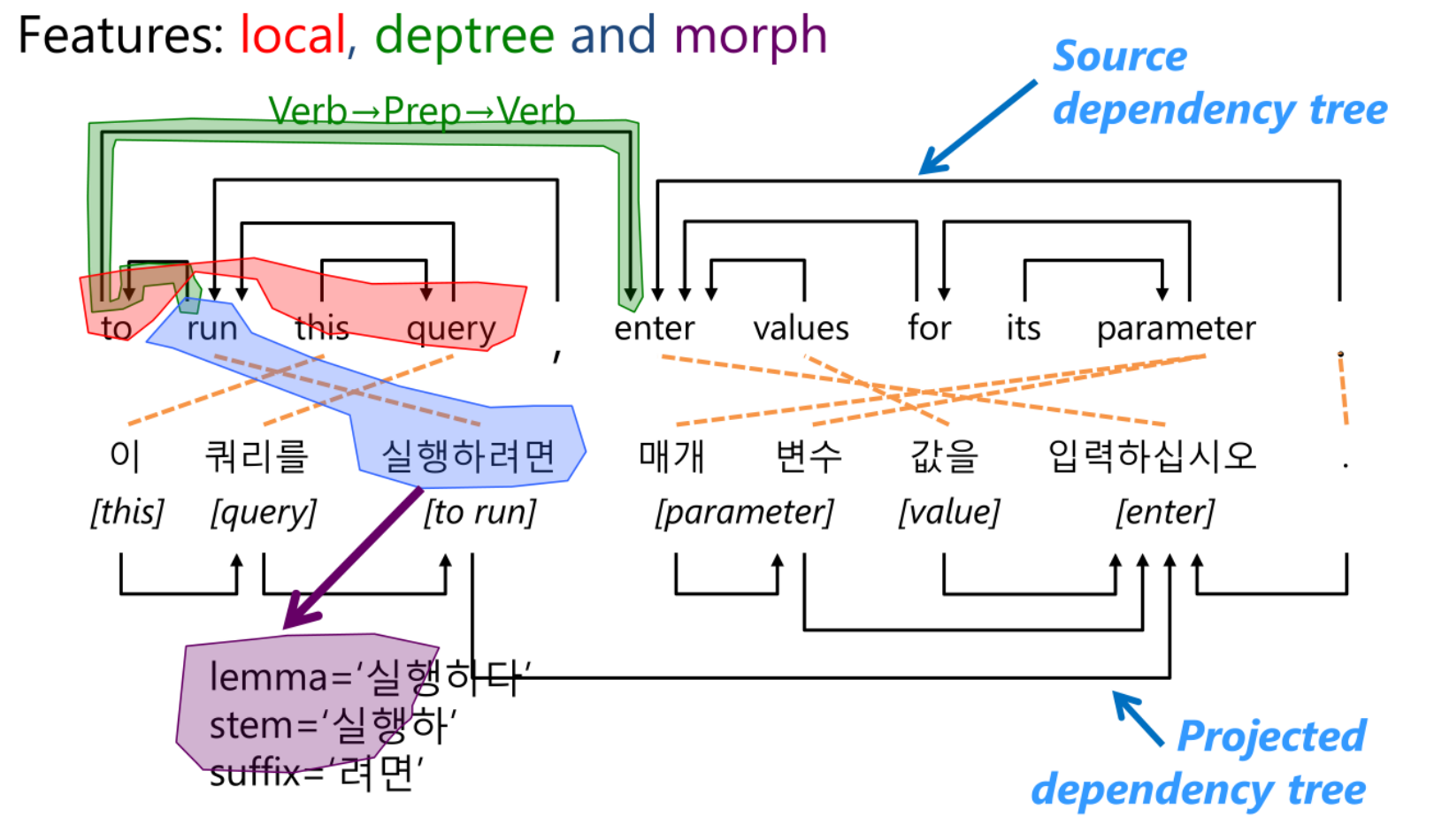
- Build a discriminative model that picks between possible translations
- Use contextual information to capture the correct translation
- Integrate into a syntax-based system with additional features

Discriminative sub-model for MT

- Given: (a) group of source words + (b) context from whole source sentence
- Predict the target translations
- Parallel data provides training pairs



Model features



Discriminative model structure

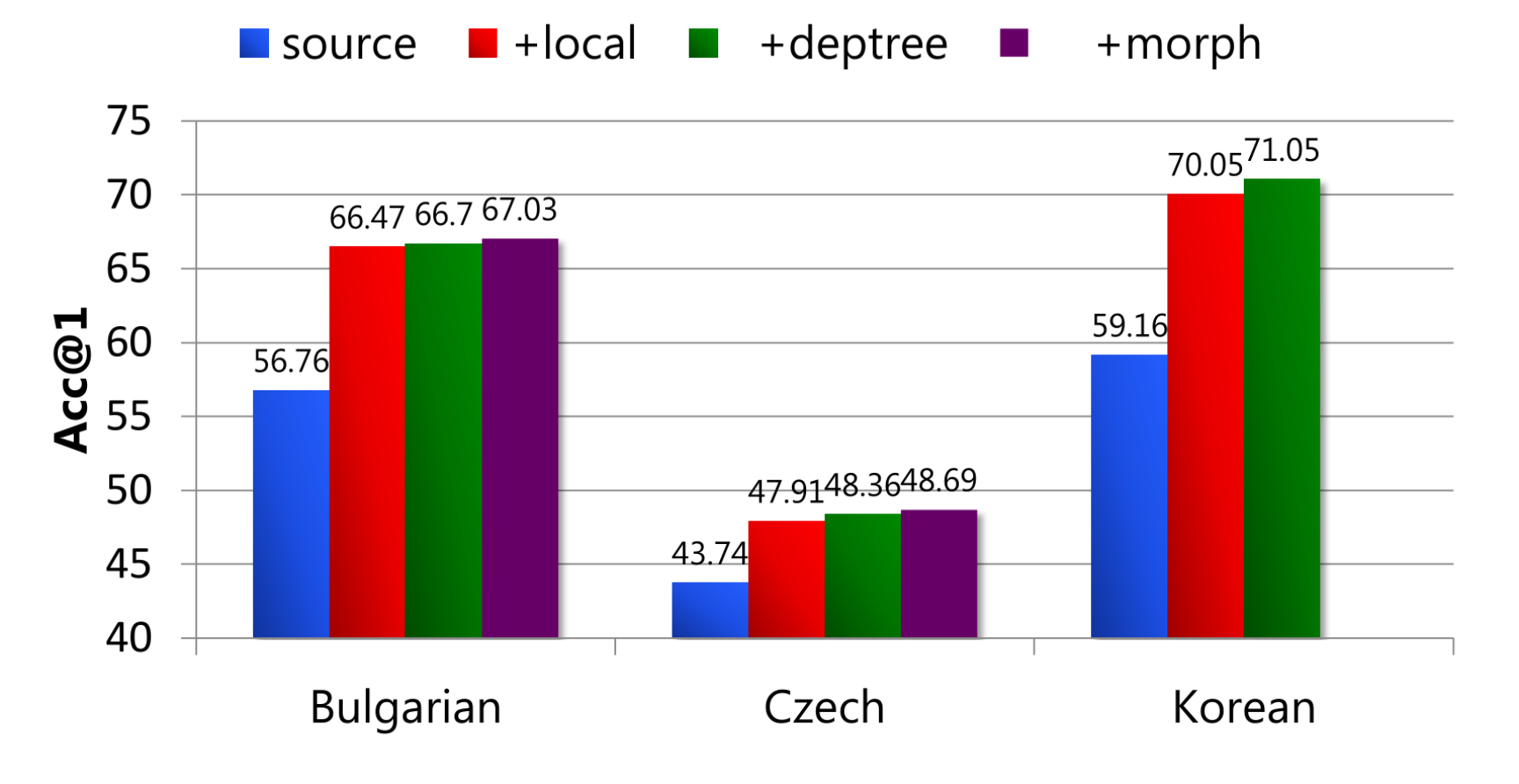
- Goal: estimate probability of target word f given $ALIGN(f)$ and $\mathbf{e}$
- Model:

$$\Pr(f|\mathbf{e}, ALIGN(f); \lambda) = \frac{\exp\left(\lambda^{\top}\phi\left(f, \mathbf{e}, ALIGN(f)\right)\right)}{\sum_{f' \in GEN(e)} \exp\left(\lambda^{\top}\phi\left(f, \mathbf{e}, ALIGN(f)\right)\right)}$$

ALIGN(f): set of source words aligned to f

GEN(e): set of possible translations of f

Results on word translation task



Integrating model inside translation

- Add three new features:
  - Log probability of target word
  - Score (unnormalized of target word)
  - Count of covered words

MT results

	MERT Dev		Test	
	Baseline	+DL	Baseline	+DL
Bulgarian	21.78	22.44	19.00	19.63
Czech	11.87	12.45	11.90	12.38
Korean	61.23	62.04	59.04	59.52

Table 8: Results (BLEU) on MT task

CONCLUSIONS

Morphology

- Morphology is an important challenge, deeply investigated in Machine Translation
- Main research lines include
  - Segmentation
  - Generation
  - Enriching models

All approaches presented here have successfully been tested in different experimental frameworks but with improvements over competitive baseline systems.

References

- UEFFING, N. AND NEY, H. 2003. Using pos information for statistical machine translation into morphologically rich languages. In 10th conference on European chapter of the Association for Computational Linguistics (EACL). Association for Computational Linguistics, Stroudsburg, PA, USA, 347354.
- HABASH, Niand SADAT, F.. Arabic Preprocessing Schemes for Statistical Machine Translation, In Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL), New York, 2006.
- VIRPIOJA, S., VAYRYNEN, J., CREUTZ, M., AND SADENIEMI, M. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In Machine Translation Summit XI. 491498
- KOEHN, P.and HOANG, H., 2007, Factored Translation Models,Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,Prague 868—876
- AVRAMIDIS, E. AND KOEHN, P. 2008. Enriching morphologically poor languages for statistical machine translation. In Conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT). Association for Computational Linguistics, Stroudsburg, PA, USA, 763-770.

- USKOREIT, J. and BRANTS, T. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *ACL-HLT.*
- GREEN, S. AND DENERO, J. 2012. A class-based agreement model for generating accurately inflected translations. In 50th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA.
- BOJAR, O. AND TAMCHYNA, A. 2011. Forms wanted: Training smt on monolingual data. In Workshop of Machine Translation and Morphologically-Rich Languages
- FORMIGA, L., HERN´ANDEZ, A., MARI˜NO, J., AND MONTE, E. 2012. Improving english to spanish out-ofdomain translations by morphology generalization and generation. In AMTA Workshop on Monolingual Machine Translation.
- MINKOV, E., TOUTANOVA, K., AND SUZUKI, H. 2007. Generating complex morphology for machine translation. In 45th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA
- JEONG, M. ,TOUTANOVA, K. ,SUZUKI,H. and QUIRK C., A Discriminative Lexicon Model for Complex Morphology, in *The Ninth Conference of the Association for Machine Translation in the Americas*, 2010

Syntax in Statistical Machine Translation

Chris Quirk

Marta R. Costa-jussà

Tutorial NAACL 2013

Overview

- Motivation
  - Examples of reordering / translation phenomena
- Synchronous context free grammar
  - Example derivations
  - ITG grammars
  - Reordering for ITG grammars
- Hierarchical phrase-based translation with Hiero
- Linguistically syntax-based translation approaches

Motivation for tree-based translation

- Phrases capture contextual translation and local reordering very well
- However, phrases are very brittle
  - "author of the book" → "本書的作者" tells us nothing about how to translate "author of the pamphlet" or "author of the play"
- Syntactic generalizations can help
  - Chinese construction "NOUN1的 NOUN2" often becomes "NOUN2 of NOUN1" in English
- Syntactically related words are not always contiguous
  - "Ich ***gab*** das Kennwort ***ein***" / "I ***entered*** the password"

Types of tree-based systems

- Formally tree-based but not using linguistic syntax
  - Can still model hierarchical nature of language
  - Can capture hierarchical reordering
  - Examples: phrase-based ITGs and Hiero
- Can use linguistic syntax on source, target, or both sides
  - Phrase structure trees, dependency trees

SYNCHRONOUS CONTEXT-FREE GRAMMARS

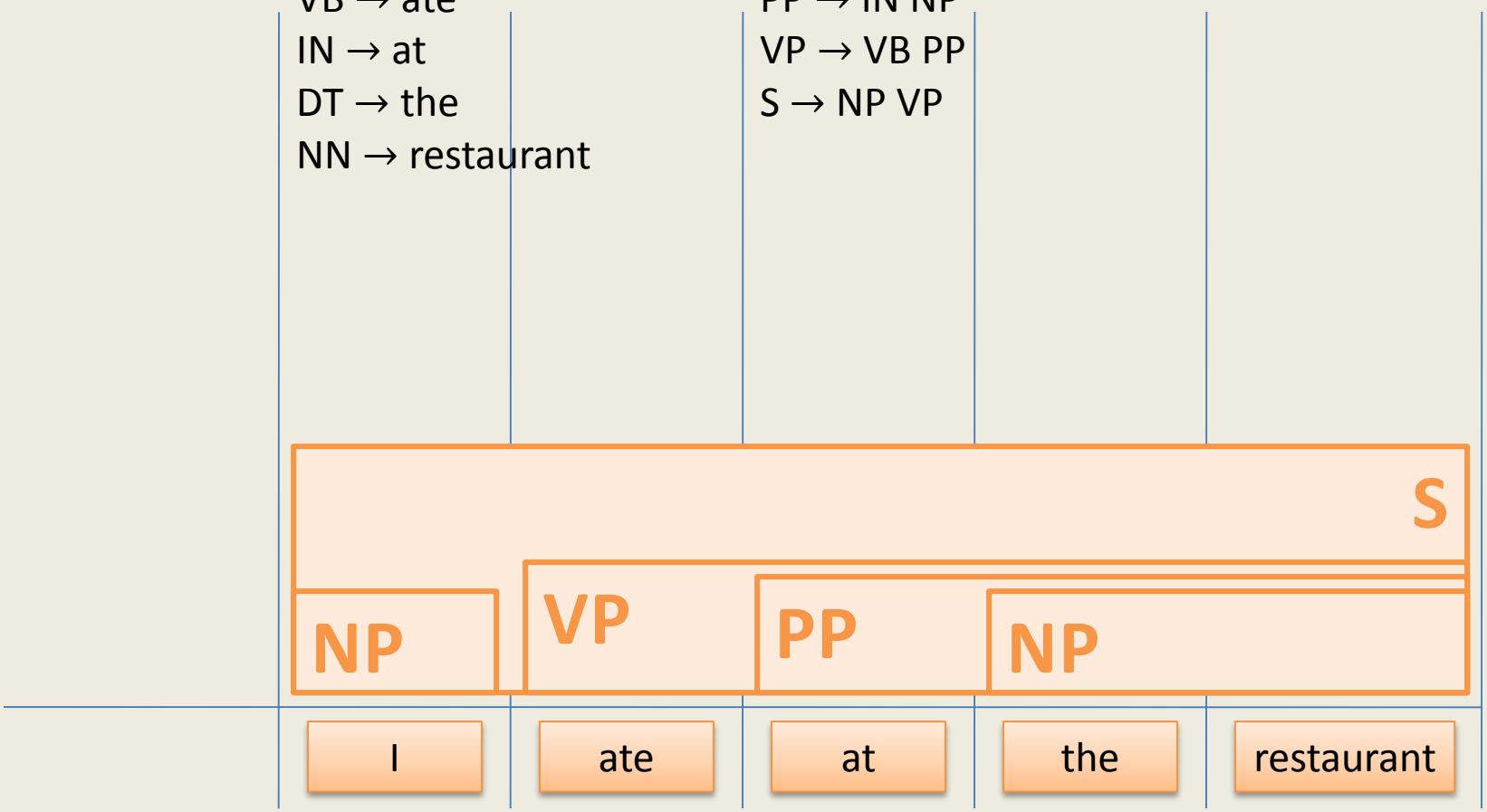
Context Free Parsing

- Grammar contains both lexical rules
  - NP→ I
  - VB → ate
  - IN → at
  - DT → the
  - NN → restaurant
- And non-lexical rules
  - NP → DT NN
  - PP → IN NP
  - VP → VB PP
  - S → NP VP
- This grammar is in Chomsky Normal Form

Context Free Parsing

NP→ I
VB → ate
IN → at
DT → the
NN → restaurant

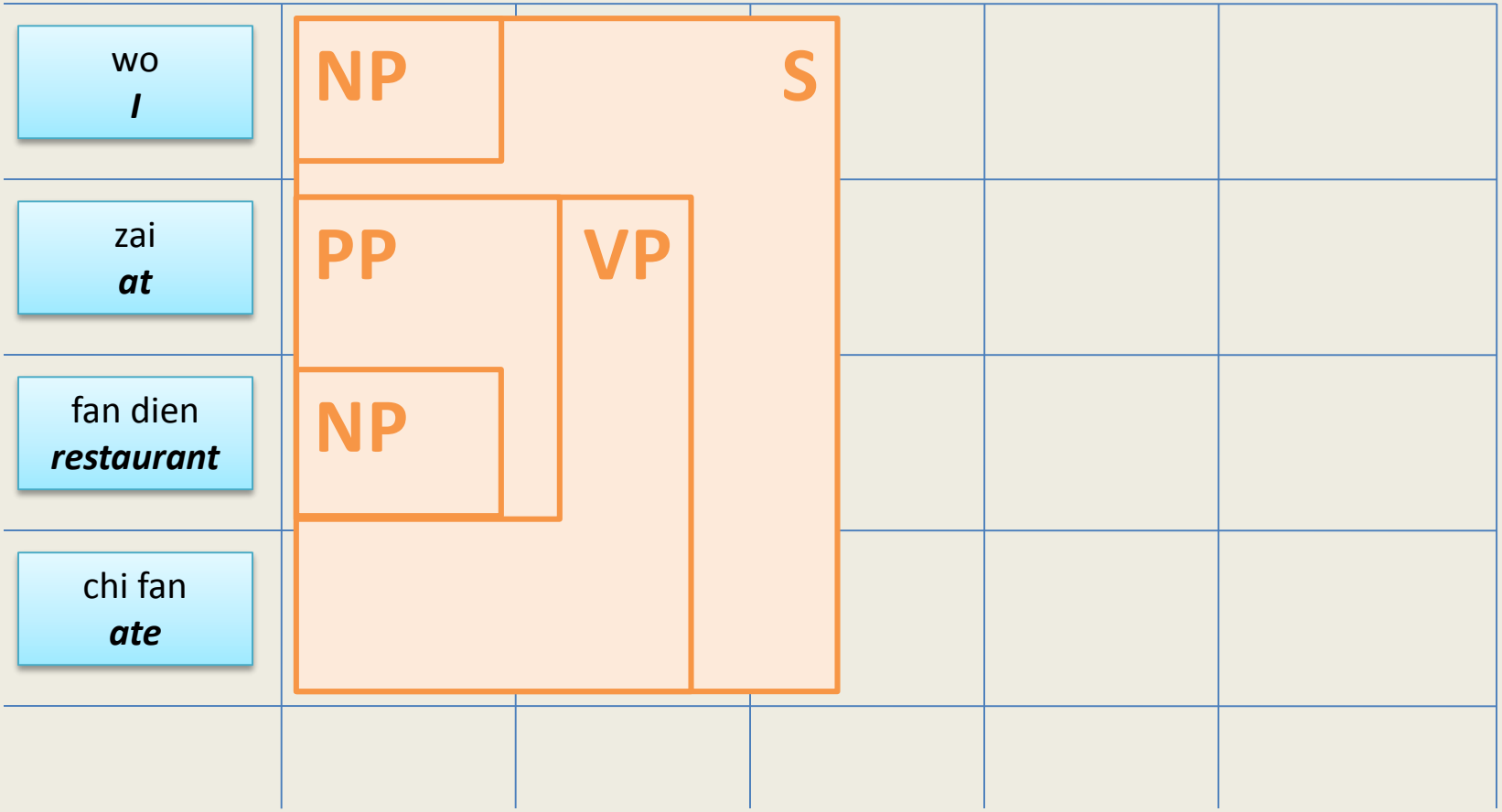
NP → DT NN
PP → IN NP
VP → VB PP
S → NP VP



Context Free Parsing

- Grammar contains both lexical rules
  - NP→ wo
  - VB → chi fan
  - IN → zai
  - NP → fan dien
- And non-lexical rules
  - NP → DT NN
  - PP → IN NP
  - VP → PP VB
  - S → NP VP

Context free parsing



Context Free Parsing

- Grammar contains both lexical rules
  - NP→ I / wo
  - VB → ate / chi fan
  - IN → at / zai
  - DT → the / ε
  - NP → restaurant / fan dien
- And non-lexical rules
  - $NP \rightarrow DT_1\ NN_2 / DT_1\ NN_2$
  - $PP \rightarrow IN_1\ NP_2 / IN_1\ NP_2$
  - $VP \rightarrow VB_1\ PP_2 / PP_2\ VB_1$
  - $S \rightarrow NP_1\ VP_2 / NP_1\ VP_2$
- Note correspondence between grammars / trees
- Can write a single grammar to parse both!

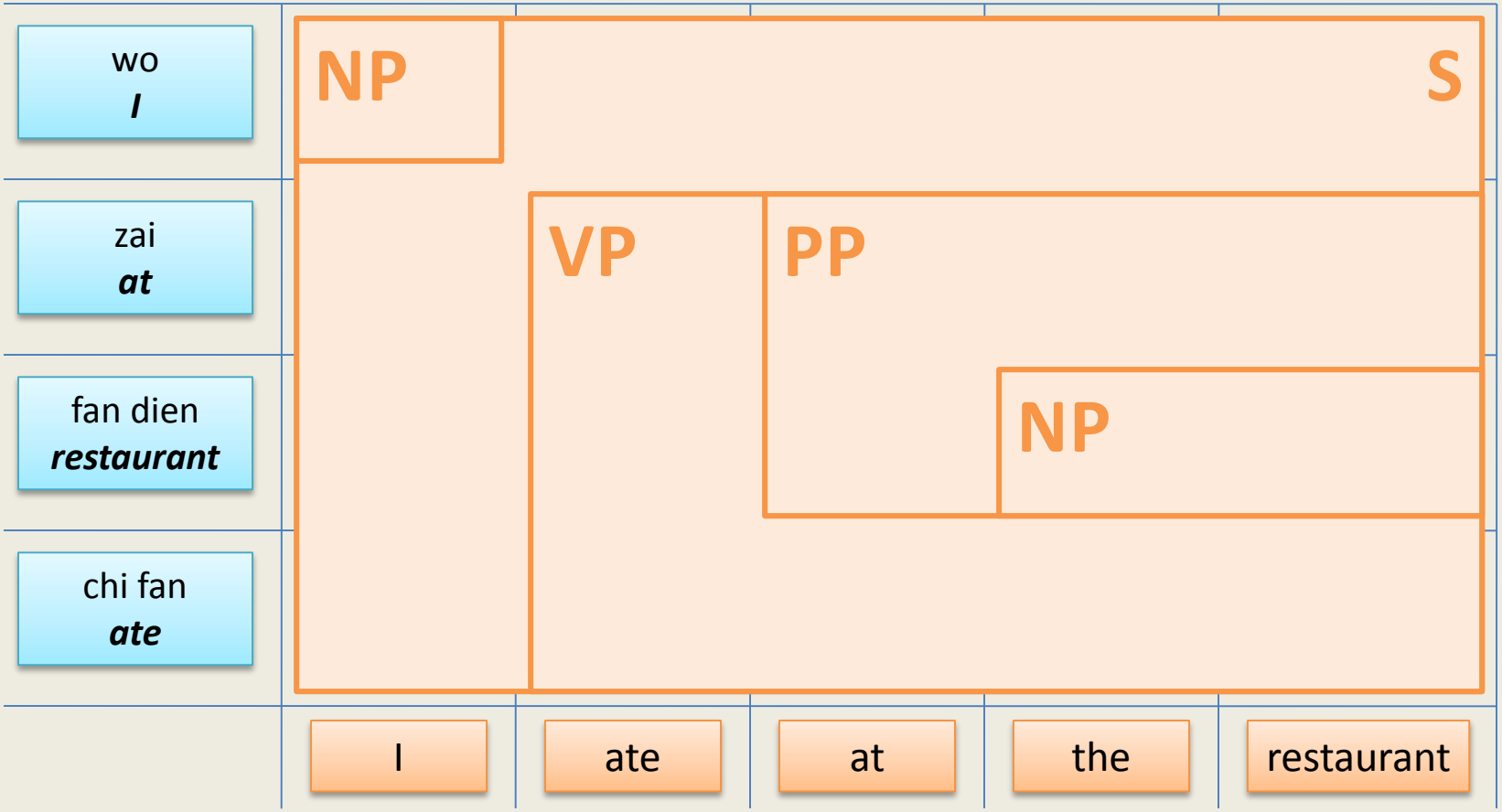
Context Free Parsing

- Grammar contains both lexical rules
  - NP→ I / wo (0.05)
  - VB → ate / chi fan (0.01)
  - IN → at / zai (0.1)
  - DT → the / ε (0.4)
  - NP → restaurant / fan dien (0.3)
- And non-lexical rules
  - $NP \rightarrow DT_1\ NN_2$ / $DT_1\ NN_2$(0.3)
  - $PP \rightarrow IN_1\ NP_2$ / $IN_1\ NP_2$(0.6)
  - $VP \rightarrow VB_1\ PP_2$ / $PP_2\ VB_1$ (0.3)
  - $S \rightarrow NP_1\ VP_2$ / $NP_1\ VP_2$ (0.4)
- Just like PCFGs, we can weight rules
- Probability of a derivation is the product of rule probabilities

Algorithms

- The same grammar can:
  - Parse parallel sentence pairs (find alignments)
  - Given a source sentence, find best tree and target translation
  - Given a target sentence, find best tree and source translation

Parallel parsing



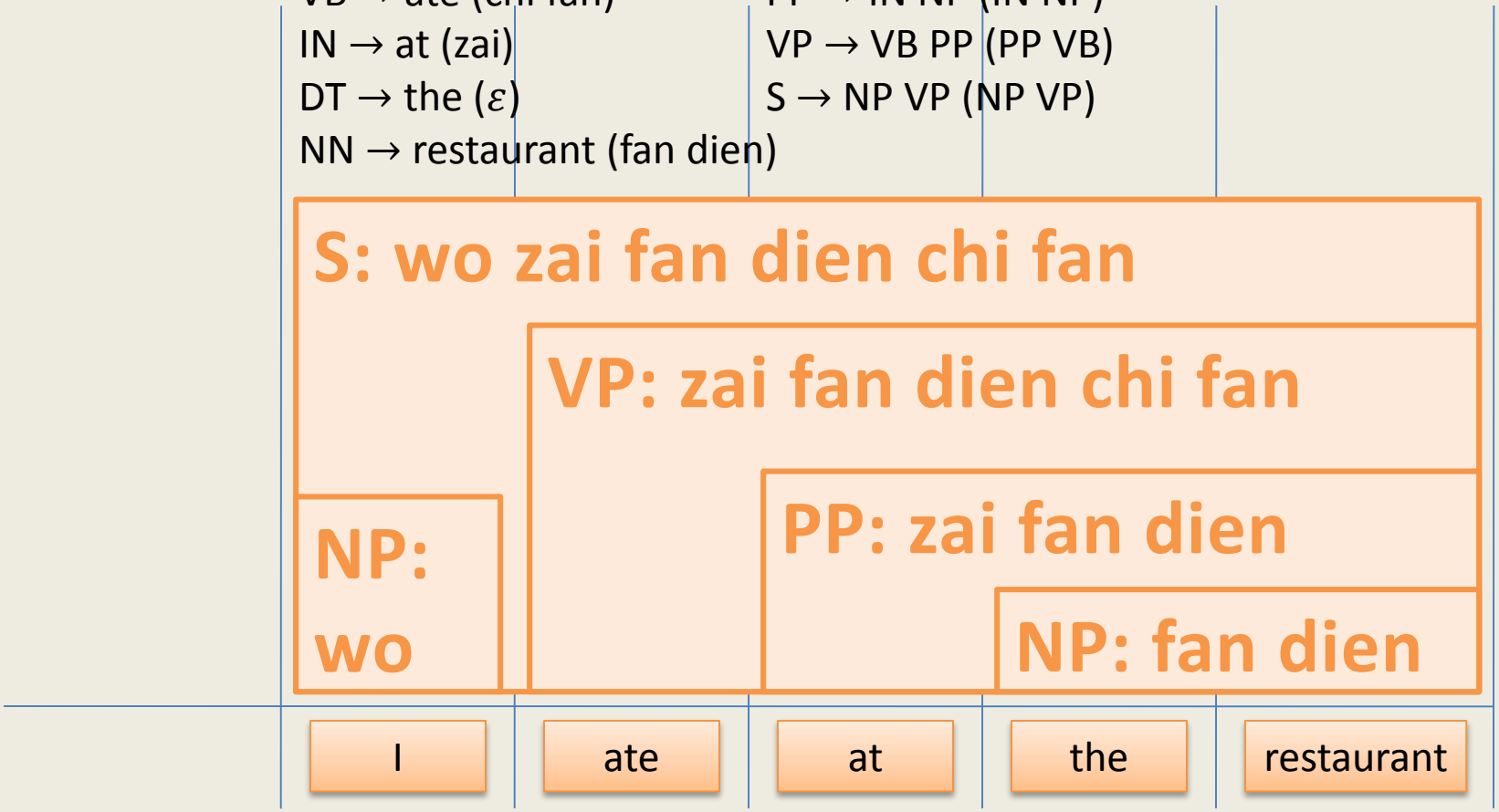
Context Free Parsing

- Grammar contains both lexical rules
  - NP→ I / wo (0.05)
  - VB → ate / chi fan (0.01)
  - IN → at / zai (0.1)
  - DT → the / ε (0.4)
  - NP → restaurant / fan dien (0.3)
- And non-lexical rules
  - $NP \rightarrow DT_1\ NN_2$ / $DT_1\ NN_2$(0.3)
  - $PP \rightarrow IN_1\ NP_2$ / $IN_1\ NP_2$(0.6)
  - $VP \rightarrow VB_1\ PP_2$ / $PP_2\ VB_1$ (0.3)
  - $S \rightarrow NP_1\ VP_2$ / $NP_1\ VP_2$ (0.4)
- Just like PCFGs, we can weight rules
- Probability of a derivation is the product of rule probabilities

Context Free Parsing

NP→ I (wo)
VB → ate (chi fan)
IN → at (zai)
DT → the (ε)
NN → restaurant (fan dien)

NP → DT NN (DT NN)
PP → IN NP (IN NP)
VP → VB PP (PP VB)
S → NP VP (NP VP)



Computational complexity

- CFG parsing
  - Runtime: $\mathcal{O}(Gn^3)$
  - Storage: $\mathcal{O}(Gn^2)$
- SCFG parsing
  - Runtime: $\mathcal{O}(Gn^6)$
  - Storage: $\mathcal{O}(Gn^4)$

- …if we have a grammar in Chomsky Normal Form!
  - All CFGs can be mapped to CNF
  - Some SCFGs may be binarized
- Grammars can be large, especially after binarization
- Pruning and approximate algorithms are important

INVERSION TRANSDUCTION GRAMMARS (ITGS)

Stochastic Inversion Transduction Grammars [Wu 97]

- A restricted form of SCFGs

$$A \rightarrow B_1 C_2, \mathrm{B_1 C_2} \text{ or } A \rightarrow [B\ C]$$

$$A \rightarrow B_1 C_2, \mathrm{C_2 B_1} \text{ or } A \rightarrow < B\ C >$$

$$A \rightarrow x, y$$

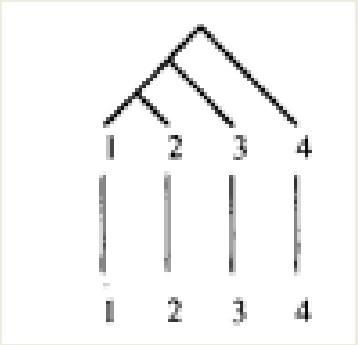
- At most binary rules, either same or reversed order
- Only non-terminals or only terminals on the right-hand-side
- This is a normal form for ITG grammars
- Bracketing grammars: only one non-terminal

Example re-ordering with ITG

Grammar includes $A \rightarrow 1,1$; $A \rightarrow 2,2$; $A \rightarrow 3,3$; $A \rightarrow 4,4$

Can the bracketing ITG generate these sentence pairs?

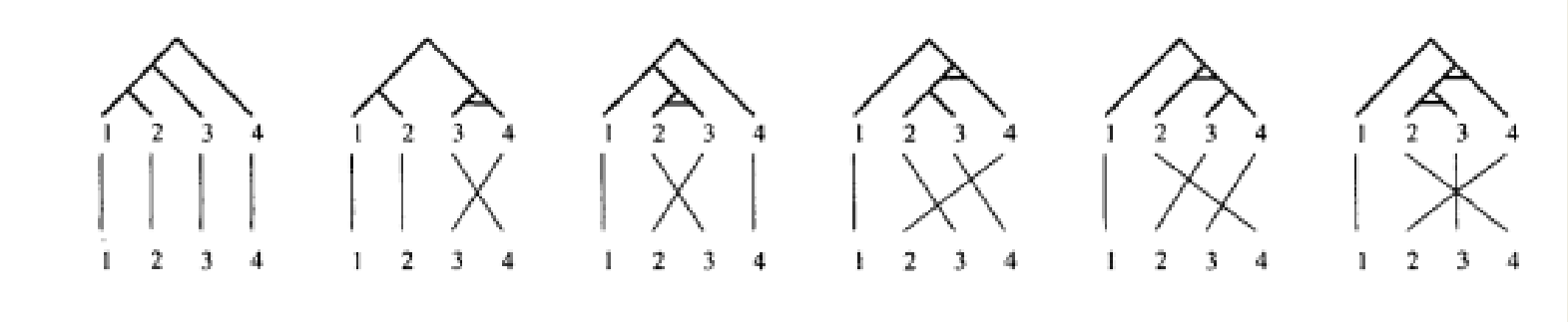
[1,2,3,4] [1,2,3,4]



$$A_1 \rightarrow [A_2 A_3]$$
$$A_2 \rightarrow [A_4 A_5]$$
$$A_4 \rightarrow [A_6 A_7]$$
$$A_6 \rightarrow 1,1$$
$$A_7 \rightarrow 2,2$$
$$A_5 \rightarrow 3,3$$
$$A_3 \rightarrow 4,4$$

Example re-ordering with ITG

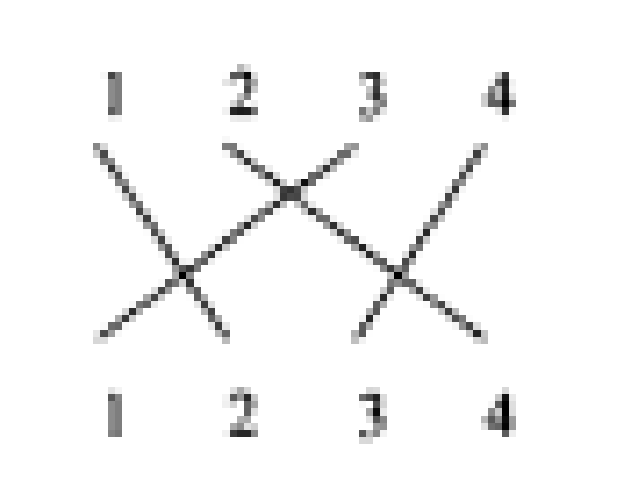
- Other re-orderings with parses

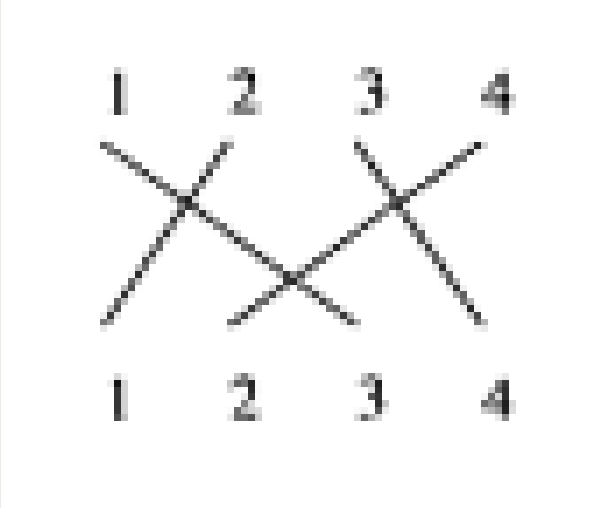


- A horizontal bar means the non-terminals are swapped

But some re-orderings are not allowed

- When words move inside-out





- 22 out of the 24 permutations of 4 words are parsable by the bracketing ITG

Number of permutations compared to ones parsable by ITG

r	ITG	all matchings	ratio
0	1	1	1.000
1	1	1	1.000
2	2	2	1.000
3	6	6	1.000
4	22	24	0.917
5	90	120	0.750
6	394	720	0.547
7	1,806	5,040	0.358
8	8,558	40,320	0.212
9	41,586	362,880	0.115
10	206,098	3,628,800	0.057
11	1,037,718	39,916,800	0.026
12	5,293,446	479,001,600	0.011
13	27,297,738	6,227,020,800	0.004
14	142,078,746	87,178,291,200	0.002
15	745,387,038	1,307,674,368,000	0.001
16	3,937,603,038	20,922,789,888,000	0.000

Application of ITGs

- Have been applied to word alignment and translation in many previous works
- Alignments can be quite good; see for instance:
  - Aria Haghighi, John Blitzer, John DeNero, and Dan Klein “Better word alignments with supervised ITG Models” 2009
- Some works in translation as well, with phrases at leaves
  - Deyi Xiong, Qun Liu, and Shouxun Lin “Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation” ACL 2006

David Chiang
ISI, USC

HIERARCHICAL PHRASE-BASED TRANSLATION

A*STAR Microsoft Research MARIE CURIE ACTIONS

Idea: Hierarchical phrases

澳洲	是	与	北韩	有	邦交	的	少数	国家	之一	。
Aozhou	shi	yu	Beihan	you	bangjiao	de	shaoshu	guojia	zhiyi	.
Australia	is	with	North Korea	have	dipl. rels.	that	few	countries	one of	.

$$yu\ X_1 you\ X_2, \qquad have\ X_2 with\ X_1$$

- The variables stand for corresponding hierarchical phrases
- Capture the fact that PP phrases tend to be before the verb in Chinese and after the verb in English
- Serves as both a discontinuous phrase pair and re-ordering rule

Other example hierarchical phrases

澳洲	是	与	北韩	有	邦交	的	少数	国家	之一	。
Aozhou	shi	yu	Beihan	you	bangjiao	de	shaoshu	guojia	zhiyi	.
Australia	is	with	North Korea	have	dipl. rels.	that	few	countries	one of	.

$$X_1 de\ X_2, \qquad the\ X_2 that\ X_1$$

- Chinese relative clauses modify NPs on the left, and English relative clauses modify NPs on the right

$$X_1 zhiyi, \qquad one\ of\ X_1$$

[Aozhou] [shi] [[[yu [Beihan]$_1$ you [bangjiao]$_2$] de [shaoshu guojia]$_3$] zhiyi]

[Australia] [is] [one of [the [few countries]$_3$ that [have [dipl. rels.]$_2$ with [N. Korea]$_1$]]]

A Synchronous CFG for example

Only 1 non-terminal X plus start symbol S used

- $X \rightarrow yu\ X_1 you\ X_2,\ have\ X_2 with\ X_1$
- $X \rightarrow X_1 de\ X_2,\ \ the\ X_2 that\ X_1$
- $X \rightarrow\ X_1 zhiyi,\ \ one\ of\ X_1$
- $X \rightarrow Aozhou,\ Australia$
- $X \rightarrow Beihan,\ North\ Korea$
- $X \rightarrow shi,\ is$
- $X \rightarrow banjiao, diplomatic\ relations$
- $X \rightarrow Aozhou,\ Australia$
- $X \rightarrow shaoshu\ guojia, few\ countries$
- $S \rightarrow S_1 X_2,\ S_1\ X_2$ [glue rule]
- $S \rightarrow X_1, X_1$

General approach

- Align parallel training data using word-alignment models (e.g. GIZA++)
- Extract hierarchical phrase pairs
  - Can be represented as SCFG rules
- Assign probabilities (scores) to rules
  - Like in log-linear models for phrase-based MT, can define various features on rules to come up with rule scores
- Translating new sentences
  - Parsing with an SCFG grammar
  - Integrating a language model

Example derivation

$\langle S_{[1]}, S_{[1]} \rangle$

$\overset{(14)}{\Rightarrow} \langle S_{[2]} X_{[3]}, S_{[2]} X_{[3]} \rangle$

$\overset{(14)}{\Rightarrow} \langle S_{[4]} X_{[5]} X_{[3]}, S_{[4]} X_{[5]} X_{[3]} \rangle$

$\overset{(15)}{\Rightarrow} \langle X_{[6]} X_{[5]} X_{[3]}, X_{[6]} X_{[5]} X_{[3]} \rangle$

$\overset{(9)}{\Rightarrow} \langle$ Aozhou $X_{[5]} X_{[3]}$, Australia $X_{[5]} X_{[3]} \rangle$

$\overset{(11)}{\Rightarrow} \langle$ Aozhou shi $X_{[3]}$, Australia is $X_{[3]} \rangle$

$\overset{(8)}{\Rightarrow} \langle$ Aozhou shi $X_{[7]}$ zhiyi, Australia is one of $X_{[7]} \rangle$

$\overset{(7)}{\Rightarrow} \langle$ Aozhou shi $X_{[8]}$ de $X_{[9]}$ zhiyi, Australia is one of the $X_{[9]}$ that $X_{[8]} \rangle$

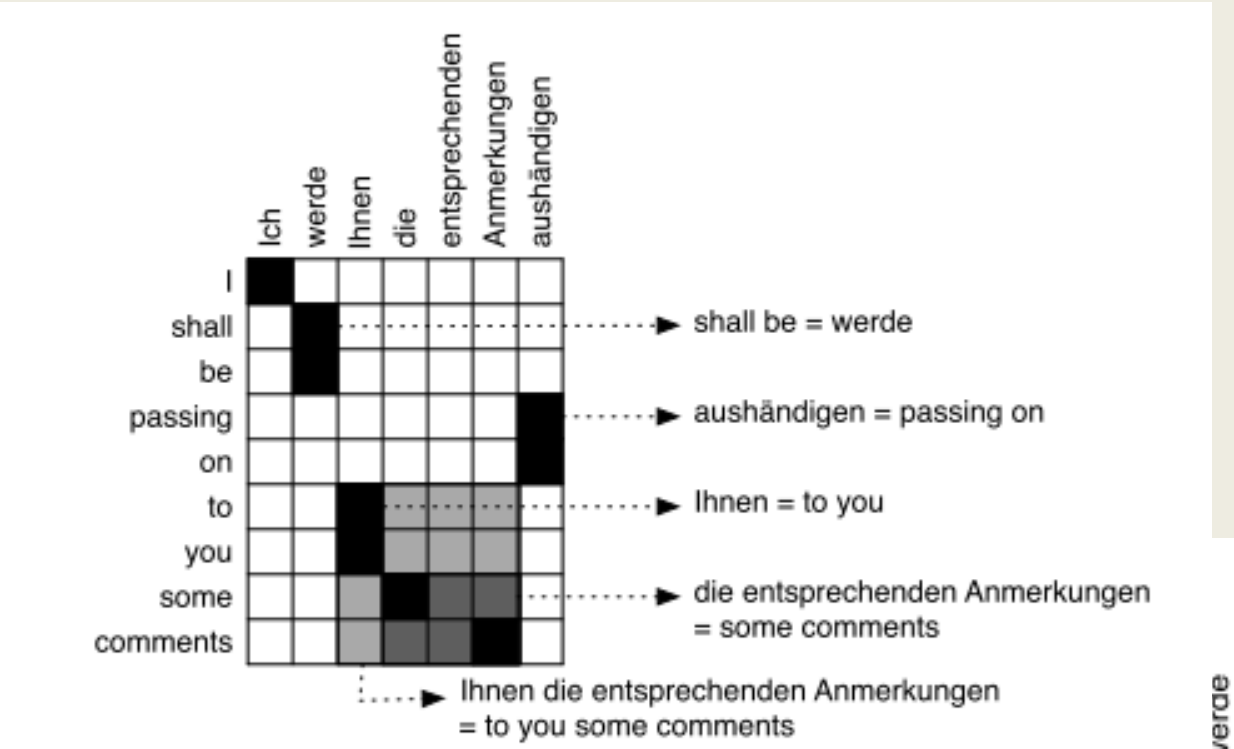
$\overset{(6)}{\Rightarrow} \langle$ Aozhou shi yu $X_{[1]}$ you $X_{[2]}$ de $X_{[9]}$ zhiyi,
Australia is one of the $X_{[9]}$ that have $X_{[2]}$ with $X_{[1]} \rangle$

$\overset{(10)}{\Rightarrow} \langle$ Aozhou shi yu Beihan you $X_{[2]}$ de $X_{[9]}$ zhiyi,
Australia is one of the $X_{[9]}$ that have $X_{[2]}$ with North Korea$\rangle$

$\overset{(12)}{\Rightarrow} \langle$ Aozhou shi yu Beihan you bangjiao de $X_{[9]}$ zhiyi,
Australia is one of the $X_{[9]}$ that have diplomatic relations with North Korea$\rangle$

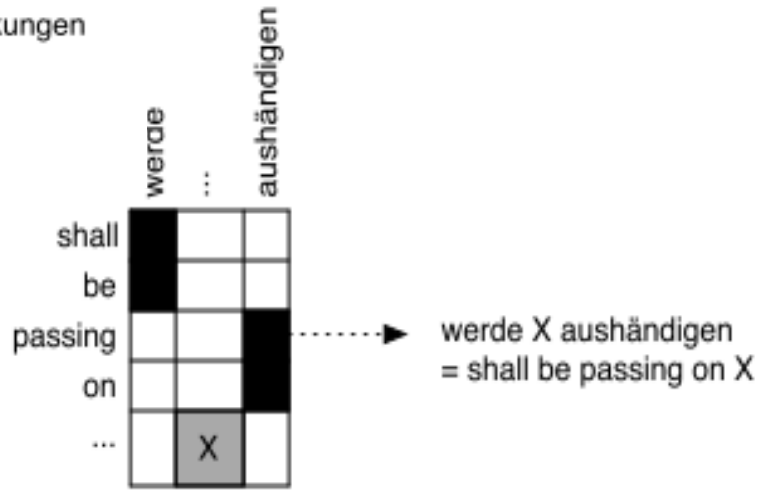
$\overset{(13)}{\Rightarrow} \langle$ Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi,
Australia is one of the few countries that have diplomatic relations with North Korea$\rangle$

Extracting hierarchical rules



Traditional phrases

Hierarchical phrase



Adding glue rules

- For continuity with phrase-based models, add glue rules which can split the source into phrases and translate each
  - $S \rightarrow S_1 X_2, S_1 X_2$
  - $S \rightarrow X_{1,} X_1$
- Question: if we only have conventional phrase pairs and these two rules, what system do we have?
- Question: what do we get if we also add these rules
  - X$\rightarrow X_1 X_2, X_1 X_2$
  - X $\rightarrow X_1 X_2, X_2 X_1$

LINGUISTICALLY SYNTAX-BASED METHODS

Phenomenology

- Languages have significant typological differences:
  - Pre-modifying adjectives in English most often are translated as post-modifying adjectives in Spanish
  - English prepositions become Japanese postpositions
  - English word order distinctions become case markers in Japanese
  - Sometimes pronouns are introduced or deleted
  - Words are often complex morphological entities, which must agree in number, gender, etc.
- The formalism is often theoretically enough to capture these phenomena, but it's difficult to learn the correct parameters
- Linguistic insights can guide parameterizations

Syntax-directed translation

- Begin by parsing source sentence
  - Syntactic analysis can guide reordering and help group related words
- One approach: Treelet translation (Quirk, Menezes, and Cherry, 2005)
  - Use ***dependency trees***: minimal amount of syntactic information (just head node)

Example (2)

- Hierarchical systems (Chiang, 2005)
  - In theory could learn X1 X2 → X2 *de* X1
  - With only one non-terminal this is too general
  - Furthermore, rules must include one aligned word pair
- Linguistic approaches can learn this

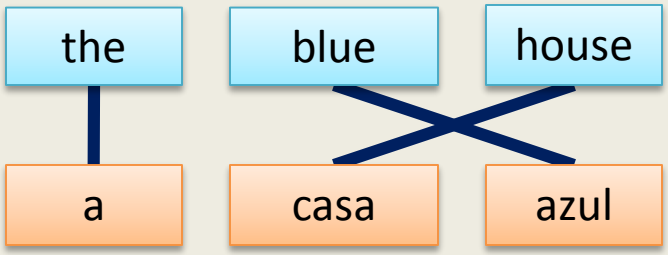
  NP(NN:x0 NN:x1) → x1 *de* x0

  However, joint modeling of lexical and syntactic transformations leads to data sparsity; can impact generalization (Wang, Knight, Marcu, 2007)

  Some of the better approaches wouldn't learn rules like this

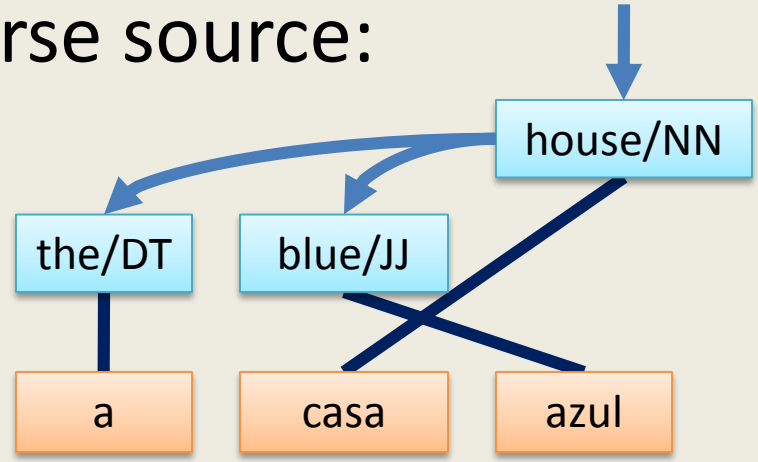
Treelet and template extraction

- Start from word aligned sentence pairs



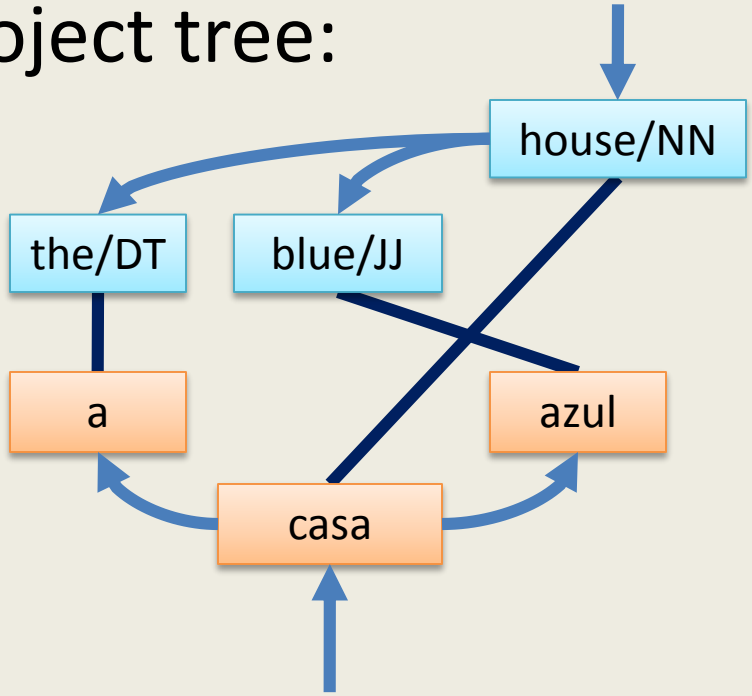
Treelet and template extraction

- Parse source:



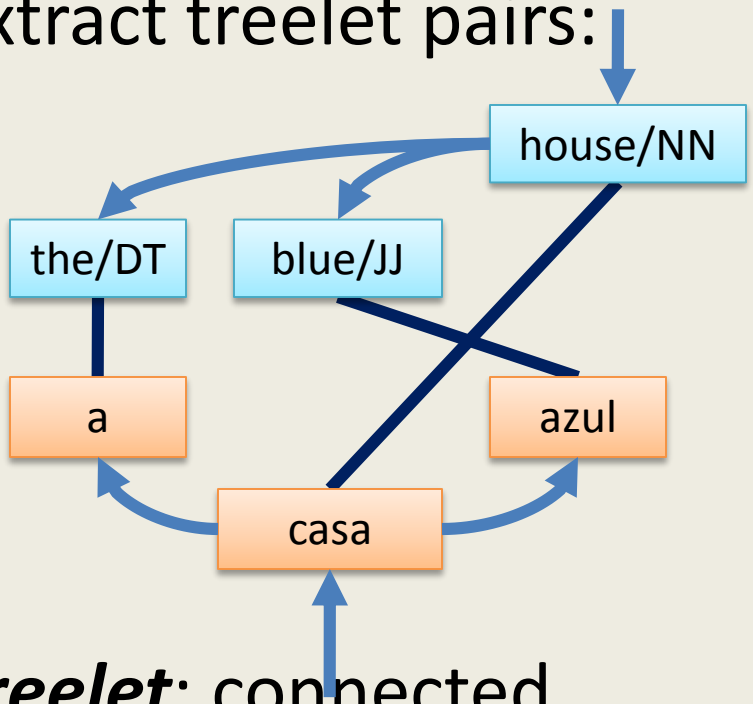
Treelet and template extraction

- Project tree:



Treelet and template extraction

- Extract treelet pairs:

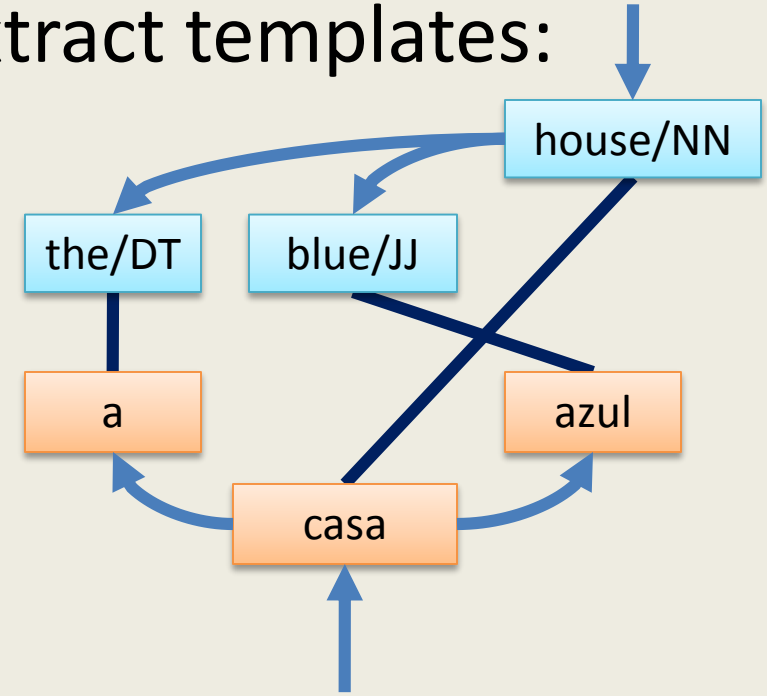


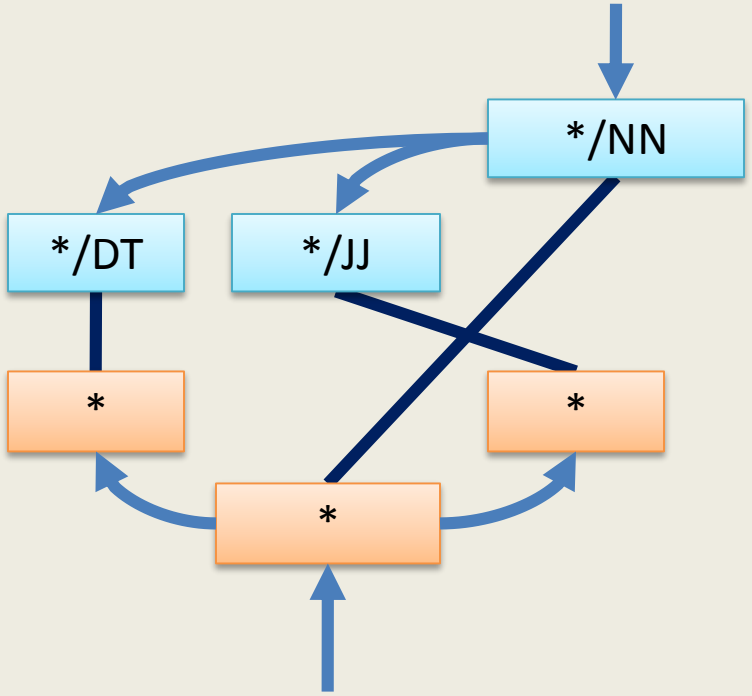
- ***Treelet***: connected subgraph of the dependency tree

- the → a
- blue → azul
- house → casa
- blue house → casa azul
- the blue house → a casa azul
- the house → a casa

Treelet and template extraction

- Extract templates:





Runtime

- Parse input sentence
- Find matching treelets and templates
- Unify to form SCFG rules
- Find best translation according to parse

Analysis of syntactic insertion and deletion Japanese → English

Insertion	Count	%	Type
の	2844	42%	Posp
を	1637	24%	Accusative
は	630	9.3%	Topic
、	517	7.6%	Punc
に	476	7.0%	Dative
する	266	3.9%	Light verb
で	101	1.5%	Posp/Copula
が	68	1.0%	Subject
して	27	0.4%	Light verb
。	26	0.4%	Punc
か	19	0.3%	Question

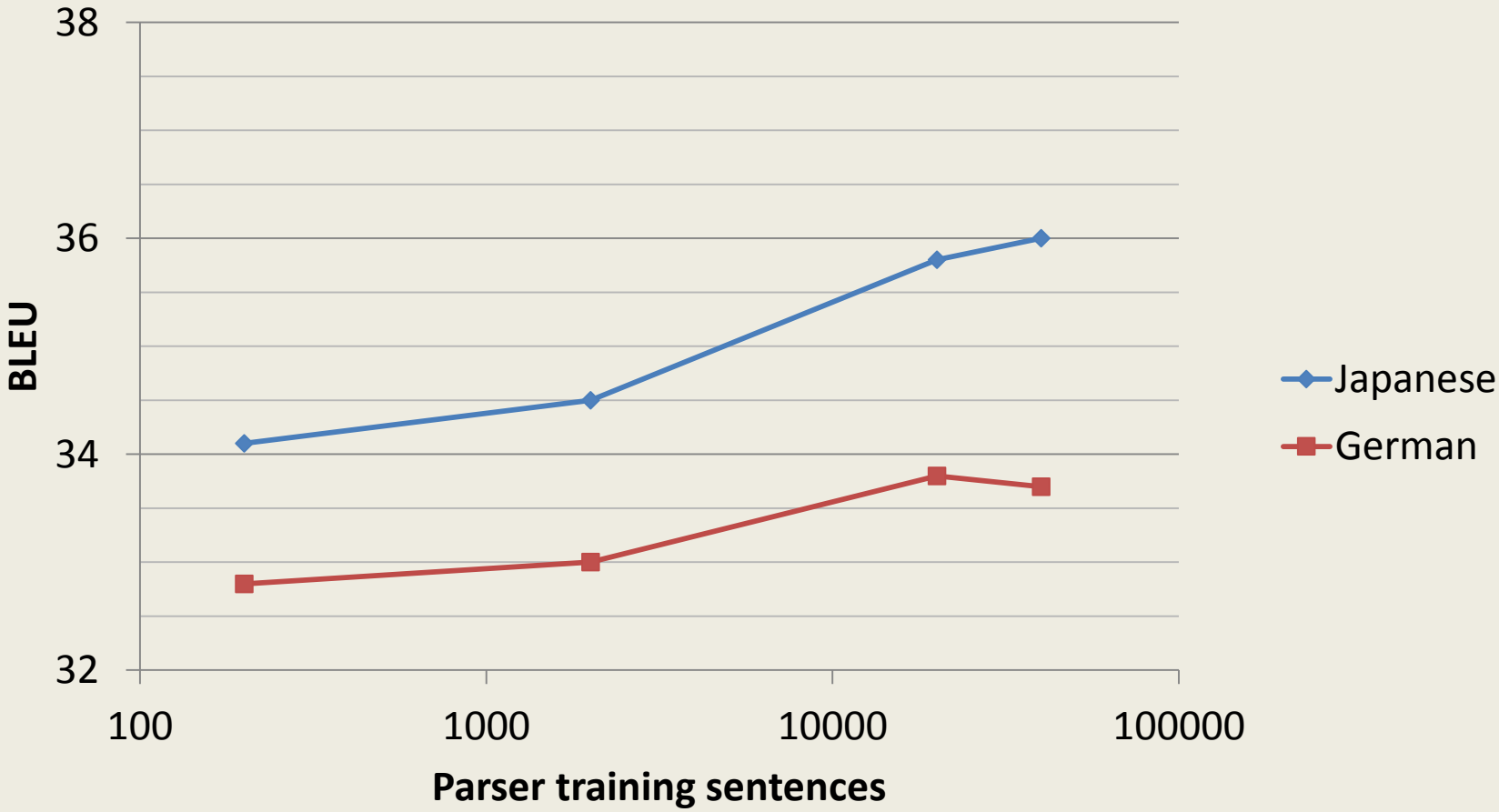
Deletion	Count	%	Type
the	875	59%	Article
-	159	11%	Punc
a	113	7.7%	Article
you	53	3.6%	Pron
it	53	3.6%	Pron
that	26	1.8%	Conj, Pron
“	23	1.6%	Punc
in	16	1.1%	Prep
.	10	0.7%	Punc
‘s	10	0.7%	Possessive
I	9	0.6%	Pron

Impact of preserving ambiguity

- Start with treelet systems
  - Technical English-German, English-Japanese
  - Newswire Chinese-English
- Translate each of k-best parses independently
- Keep the translation with the best score
- Evaluate using BLEU
- Can scale larger with packed forests

parses	EG	EJ	CE
1	33.6	36.0	28.2
2	33.8	36.1	28.5
4	34.1	36.3	28.9
8	34.3	36.6	29.2
16	34.5	36.8	29.7
32	34.8	37.1	30.0

Parse quality vs. translation quality



Target language parsing

- One of the original motivations for syntactic parsing was grammaticality
  - Goal is to identify well-formed utterances

    I went to the store

    * I the store to went
- If we want grammatical MT output, incorporate target language parsing into translation

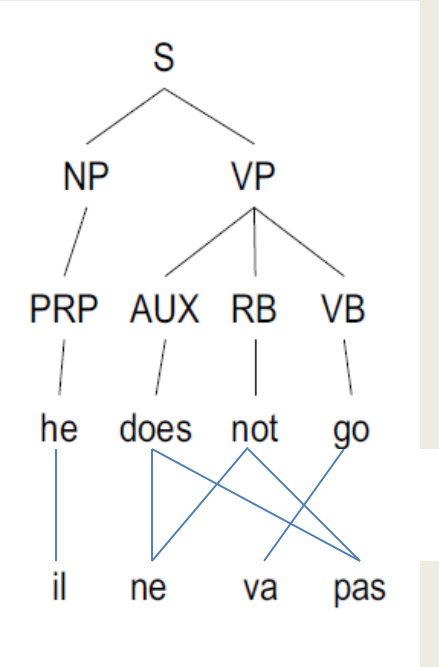
What's in a translation rule? (Galley et al. 2004)

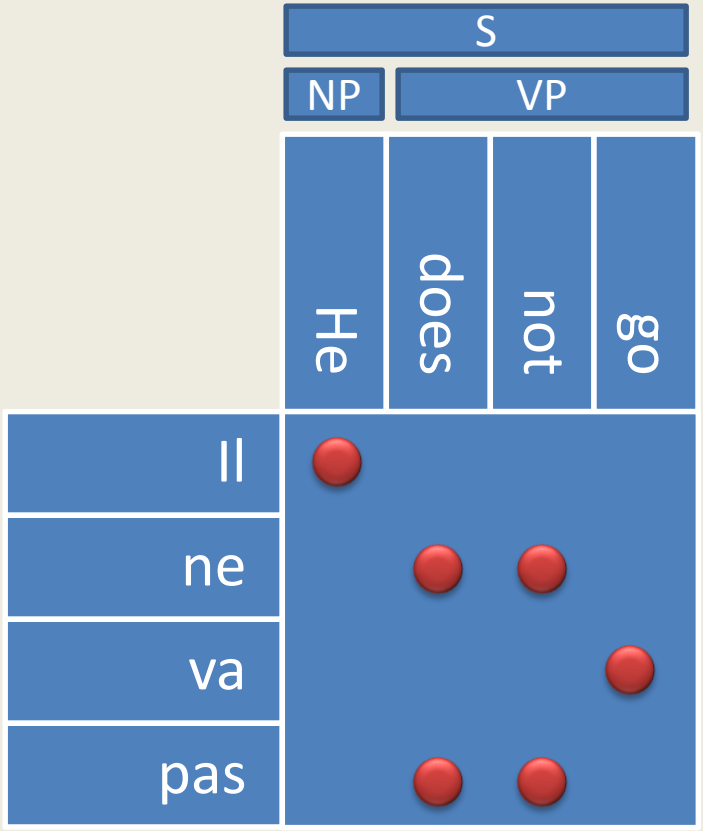
- Where do our rules acquire syntax? Why, Penn Treebank trees, of course!
  - This is not a universal choice, just common
  - It has its downsides:
    - NP structure is very flat
    - Syntactic structures aren't very detailed (for better or worse)
  - And its upsides:
    - Lots of effort in making good treebank parsers
    - Common representation for the community
- Could hand-craft all those rules – we'll try to learn them from data instead

Rule forms

- Rules so far have been single-layer synchronous
  $S \rightarrow X_1\ de\ \ X_2,\quad X_2\ \ X_1$
- We'll allow multi-layer fragments too
  $$S \rightarrow (X\ (A_1\ B_2)\ C_3),\quad C_3\ de\ A_1\ B_2$$
- This can increase the capacity of the model
- Also, we can write these as rules with variables:
  $$\big(S\ (A\ (\alpha{:}\,A\ \ \beta{:}\,B)\ \gamma{:}\,C)\big) \rightarrow \gamma\ de\ \alpha\ \beta$$
- This reflects a view of the rules as tree-to-string / string-to-tree transduction rules

Sample sentence pair





Derivations

- We'll apply a series of rewrites to transform the string in one language into a tree in another
  - il va

    (NP (PRP he)) → il /// NP → he, il
  - (NP (PRP he)) va

    (VP (VB goes)) → va /// VP → goes, va
  - (NP (PRP he)) (VP (VB goes))

    (S (α:NP) (β:VP)) → α β /// S → NP_1 VP_2, NP_1 VP_2
  - (S (NP (PRP he)) (VP (VB goes)))

Rule sets

- Given a source string S, a target tree T, and an alignment A
- Define $\rho_A(S, T)$, the set of rules extractable from all derivations $D \in \delta_A(S, T)$
- Turns out they fall into a nice lattice: there's a set of non-overlapping minimal rules, from which larger rules can be composed
- Another view: there is a set of frontier nodes $\mathcal{F}(T) \subseteq T$ which are the starting and ending points of minimal rules

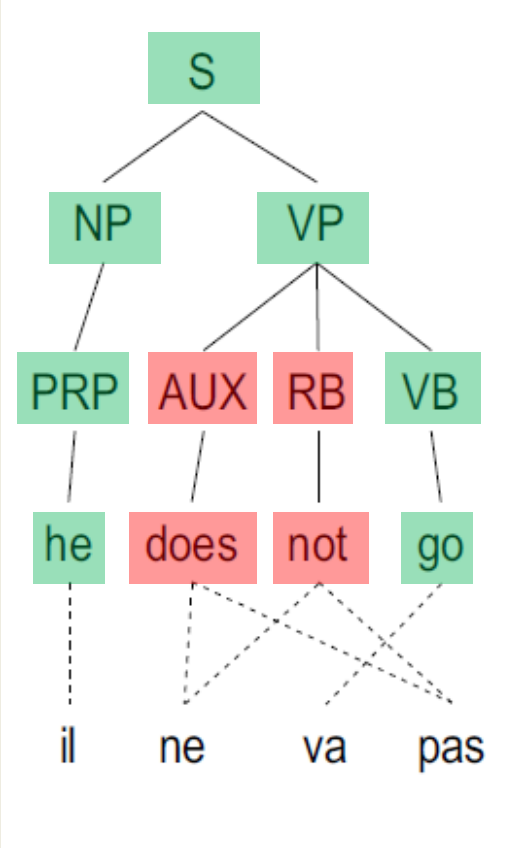
Finding frontier nodes

- A node is a ***frontier node*** if

  the range of foreign words covered inside its span

  does not overlap with

  the range of foreign words covered outside its span



Coverage vs. rule size

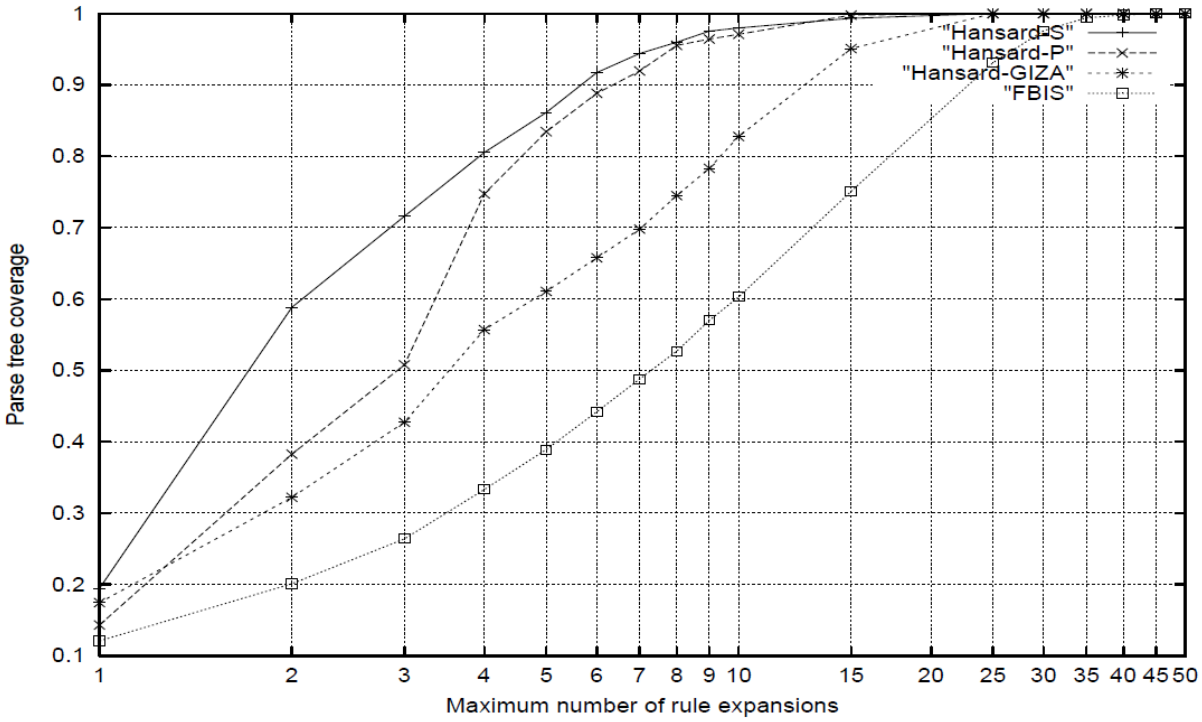


Figure 9: Percentage of parse trees covered by the model given different constraints on the maximum size of the transformation rules.

Composed rules

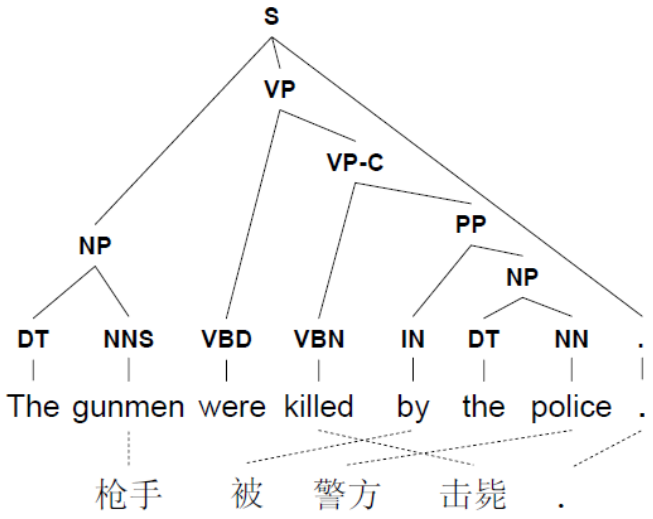
- Stepping from frontier node to frontier node gets us the bare necessities
- As in phrase-based translation, though, keeping larger fragments makes a big difference
- Numbers from a simple system: just one feature (second prob estimate), starting from minimal rules, then moving up to rules of size 3 and 4

	C_m	C_3	C_4
Chinese-to-English	24.47	27.42	28.1

Table 6: BLEU-4 scores for the 2002 NIST test set, with rules of increasing sizes.

Composed vs. minimal derivations

	Best minimal-rule derivation (C_m)	$p(r)$
(a)	S(x_0:NP-C x_1:VP x_2:.) $\rightarrow$ x_0 x_1 x_2	.845
(b)	NP-C(x_0:NPB) $\rightarrow$ x_0	.82
(c)	NPB(DT(*the*) x_0:NNS) $\rightarrow$ x_0	.507
(d)	NNS(*gunmen*) $\rightarrow$ 枪手	.559
(e)	VP(VBD(*were*) x_0:VP-C) $\rightarrow$ x_0	.434
(f)	VP-C(x_0:VBN x_1:PP) $\rightarrow$ x_1 x_0	.374
(g)	PP(x_0:IN x_1:NP-C) $\rightarrow$ x_0 x_1	.64
(h)	IN(*by*) $\rightarrow$ 被	.0067
(i)	NP-C(x_0:NPB) $\rightarrow$ x_0	.82
(j)	NPB(DT(*the*) x_0:NN) $\rightarrow$ x_0	.586
(k)	NN(*police*) $\rightarrow$ 警方	.0429
(l)	VBN(*killed*) $\rightarrow$ 击毙	.0072
(m)	.(.) $\rightarrow$.	.981



	Best composed-rule derivation (C_4)	$p(r)$
(o)	S(NP-C(NPB(DT(*the*) NNS(*gunmen*))) x_0:VP .(.)) $\rightarrow$ 枪手 x_0 .	1
(p)	VP(VBD(*were*) VP-C(x_0:VBN PP(IN(*by*) x_1:NP-C))) $\rightarrow$ 被 x_1 x_0	0.00724
(q)	NP-C(NPB(DT(*the*) NN(*police*))) $\rightarrow$ 警方	0.173
(r)	VBN(*killed*) $\rightarrow$ 击毙	0.00719

Figure 4: Two most probable derivations for the graph on the right: the top table restricted to minimal rules; the bottom one, much more probable, using a large set of composed rules. Note: the derivations are constrained on the $(\pi, \mathbf{f}, \mathbf{a})$ triple, and thus include some non-literal translations with relatively low probabilities (e.g. *killed*, which is more commonly translated as 死亡).

Lots of important issues

- Binarization can substantially improve performance (lower search error)
  - Greedy binarization approaches can work well
- Can model tree sequences instead of trees
  - Helps handle correlated adjacent phenomena
- Can “fuzz” the tree – search over “almost constituents”
  - Hao Zhang, Licheng Fang, Peng Xu, Xiaoyun Wu; Binarized Forest to String Translation, ACL 2011

References

- *Hierarchical phrase-based translation*. David Chiang, CL 2007.
- *An introduction to Synchronous Grammars.* Notes and slides from ACL 2006 tutorial. David Chiang.
- *Stochastic inversion transduction grammars and bilingual parsing of parallel corpora.* Dekai Wu, CL 1997.
- *Better word alignment with Supervised ITG models.* ACL 2009, A. Haghighi, J. Blitzer, J. DeNero, and D. Klein
- Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. ***What's in a translation rule?*** NAACL 2004.
- Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, Ignacio Thayer. ***Scalable Inference and Training of Context-Rich Syntactic Translation Models***. ACL 2006.
- Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. SPMT: Statistical ***Machine Translation with Syntactified Target Language Phrases***. EMNLP 2006.
- Liang Huang, Kevin Knight, and Aravind Joshi. ***Statistical Syntax-Directed Translation with Extended Domain of Locality***. AMTA 2006.
- Haitao Mi, Liang Huang, and Qun Liu. ***Forest-based Translation***. ACL 2008.
- Arul Menezes and Chris Quirk. ***Syntactic Models for Structural Word Insertion and Deletion during Translation***. EMNLP 2008.

Semantics in Statistical Machine Translation

Chris Quirk

Marta R. Costa-jussà

Tutorial NAACL 2013

Semantics: the study of meaning

- Often we start with ***lexical semantics***, the meanings of words
  - To translate a word correctly, we need to know what it means
- Language generally follows the ***principle of compositionality***
  - Meaning of a complex expression is a function of its parts
  - To translate a sentence correctly, we need to understand the objects and their relationships

Vauquois triangle

General idea:
The deeper our representation of language, the easier the translation task.

Flip side:
Deep analyses require complex analyzers and complex generation

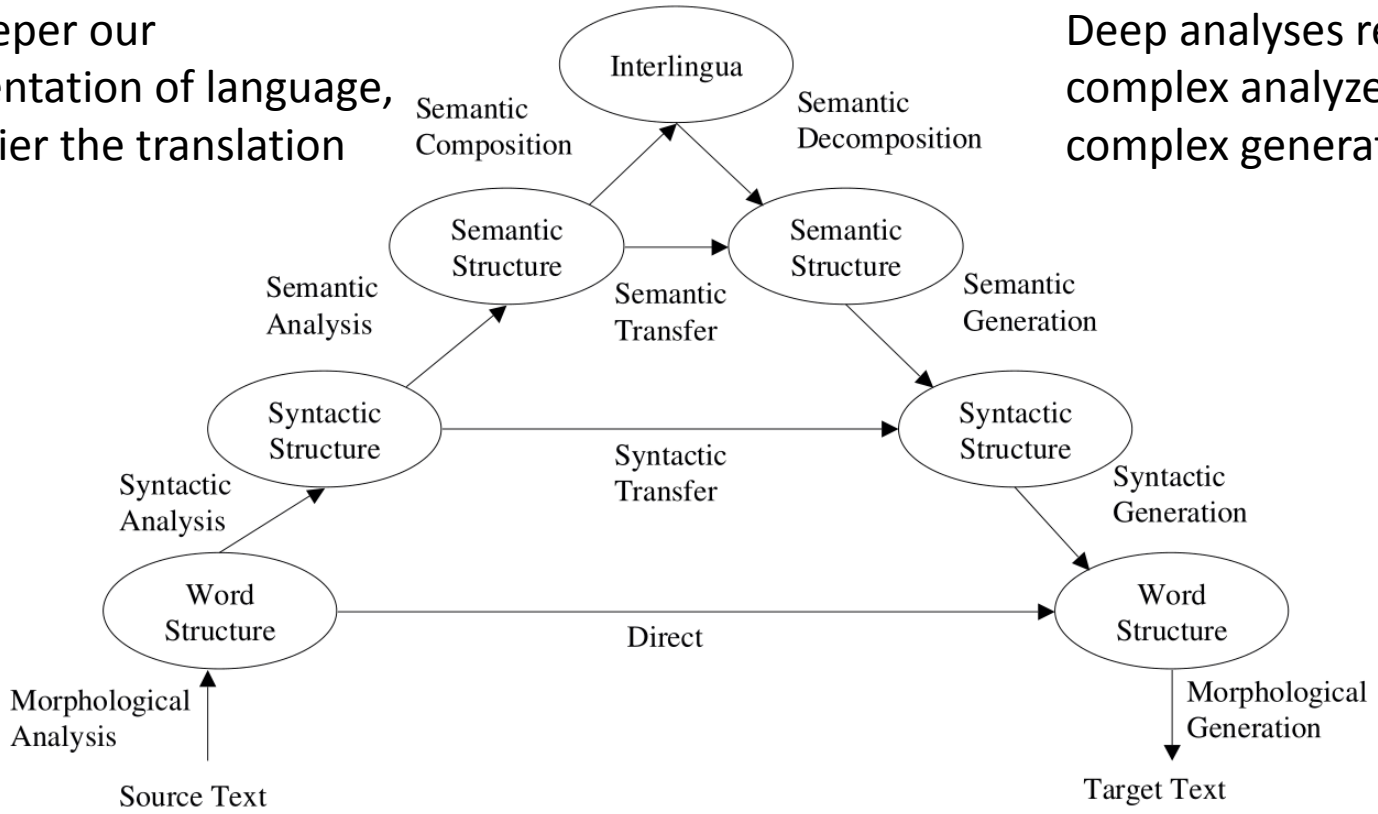


Figure 1: The Vauquois Triangle for MT

Lexical Semantics

Homonymy: same word with **multitude of unrelated meanings** (bank)

Synonymy: Different words may have **identical or similar meanings** (big, large)

Hypernymy: more general word (colour)

Holonymy: term denoting the whole (body)

Polysemy: same **word with multitude of related meanings** (man)

Antonymy: Different words **with opposite meanings** (big, small)

Hyponymy: less general word (red)

Meronymy: term denoting part of hand)

Main problem with Lexical Semantics faced in the field of Machine Translation

Homonymy: same word with **multitude of meanings** (bank)

Bank

1. Where you sit
2. Where you put the money

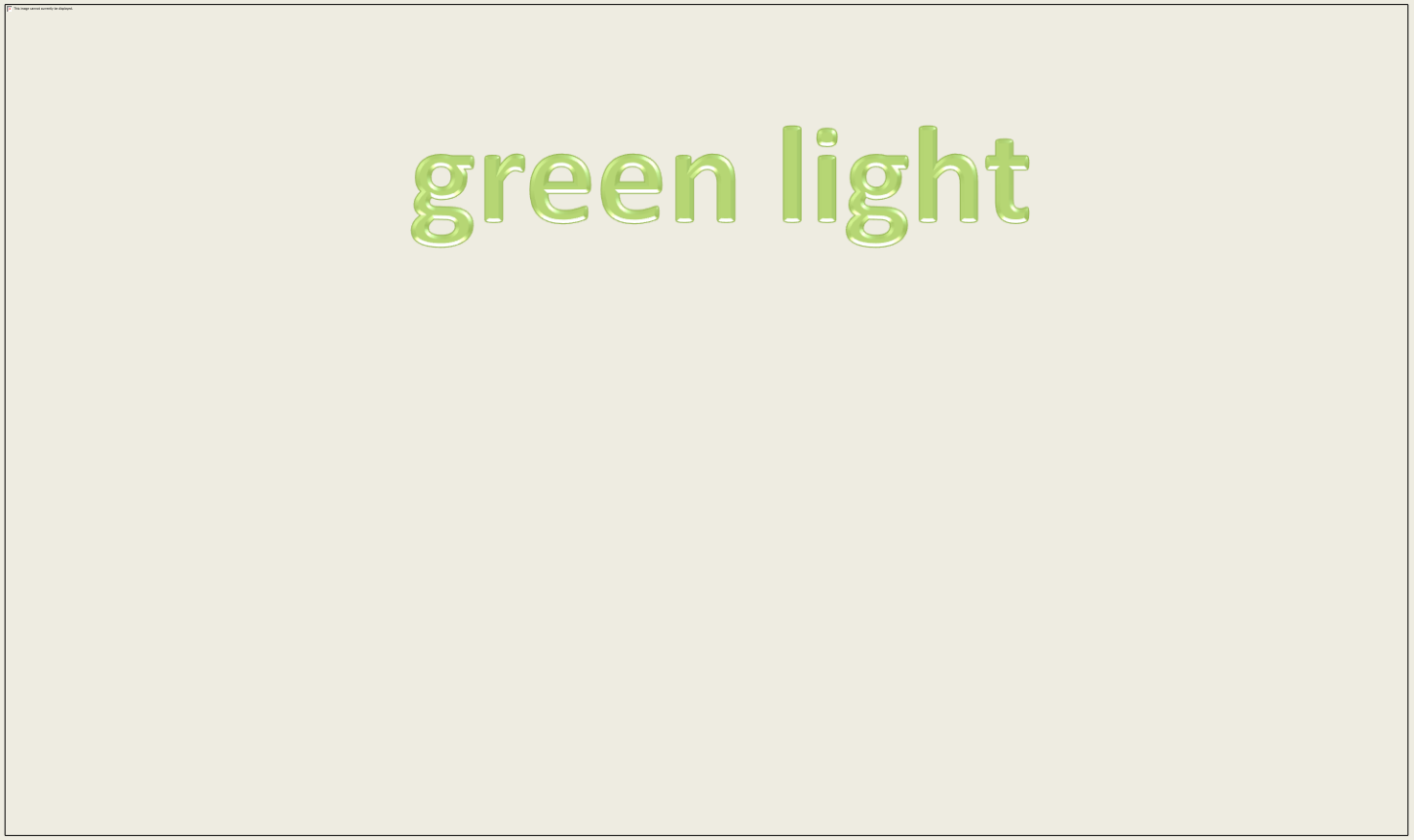
WSD from source to target

Polysemy: same **word with multitude of related meanings** (man)

Man

1. The human species (man vs animal)
2. Males of human species (man vs woman)
3. Adult males of the human species (man vs boy)

What do you think when you see the words?



Sense disambiguation

Context-dependent translation

Syntactic / semantic roles

Toward full semantic translation

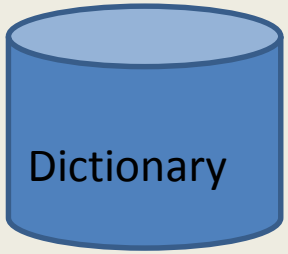
OUTLINE

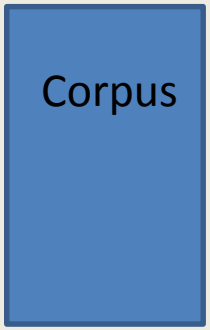
the ability to computationally determine which sense (word/phrase) is activated by its use in a particular context

SENSE DISAMBIGUATION

Disambiguation elements (WSD)

- Dictionary + Corpus





Play

1. (N) a drama
2. (V) take part in a game
3. (V) perform an instrument

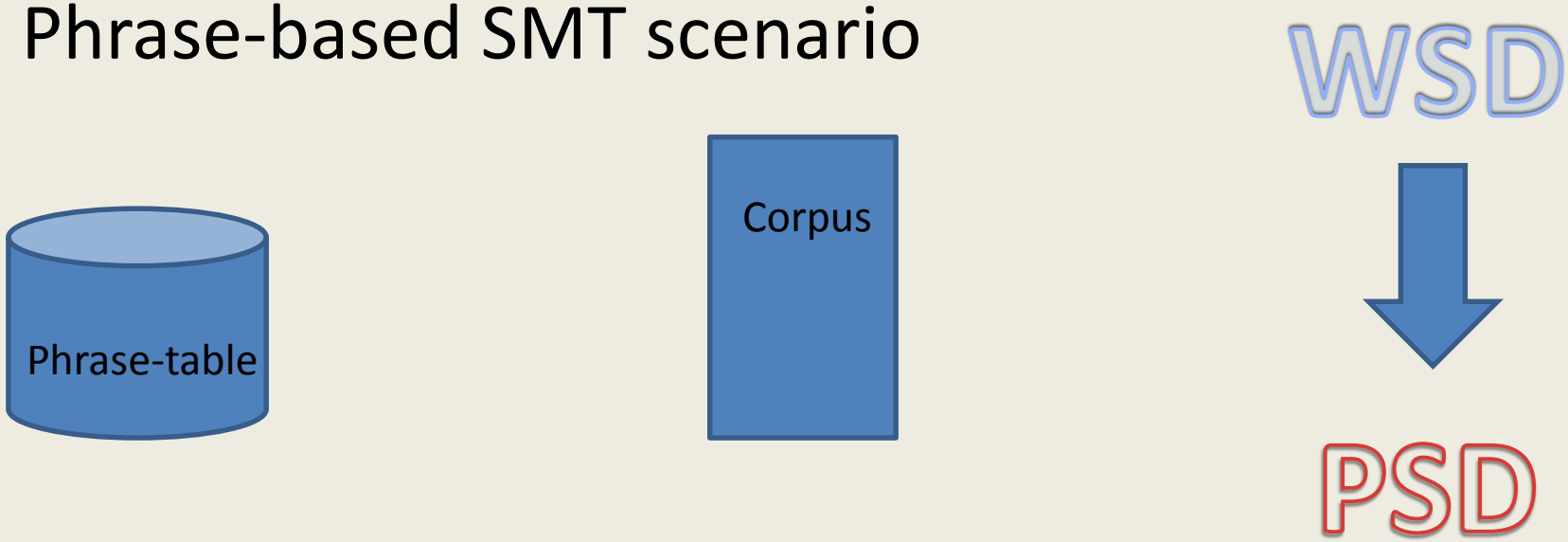
1. She was the main actress in that famous play
2. His brother did not like to play basketball
3. I love to play the piano

Sense disambiguation = classification

- word/phrase senses are the *classes, and*
- *an automatic classification method is used to assign each occurrence of a word/phrase to one or more* classes
- based on the evidence from the *context and from external knowledge sources*

Disambiguation elements (PBSMT)

- Phrase-based SMT scenario



  - The phrase-table becomes the dictionary
  - The corpus is the source text to be translated

CONTEXT-DEPENDENT TRANSLATION

Source context information

- While translating a word / phrase in a sentence.

We must consider the contexts in which that word / phrase appears.

Context-aware Translation Unit

- A new phrase (translation unit) definition is required:

  source context : source side : target side

- We would need to estimate conditional probabilities of the form:

  P(target side | source side, source context)

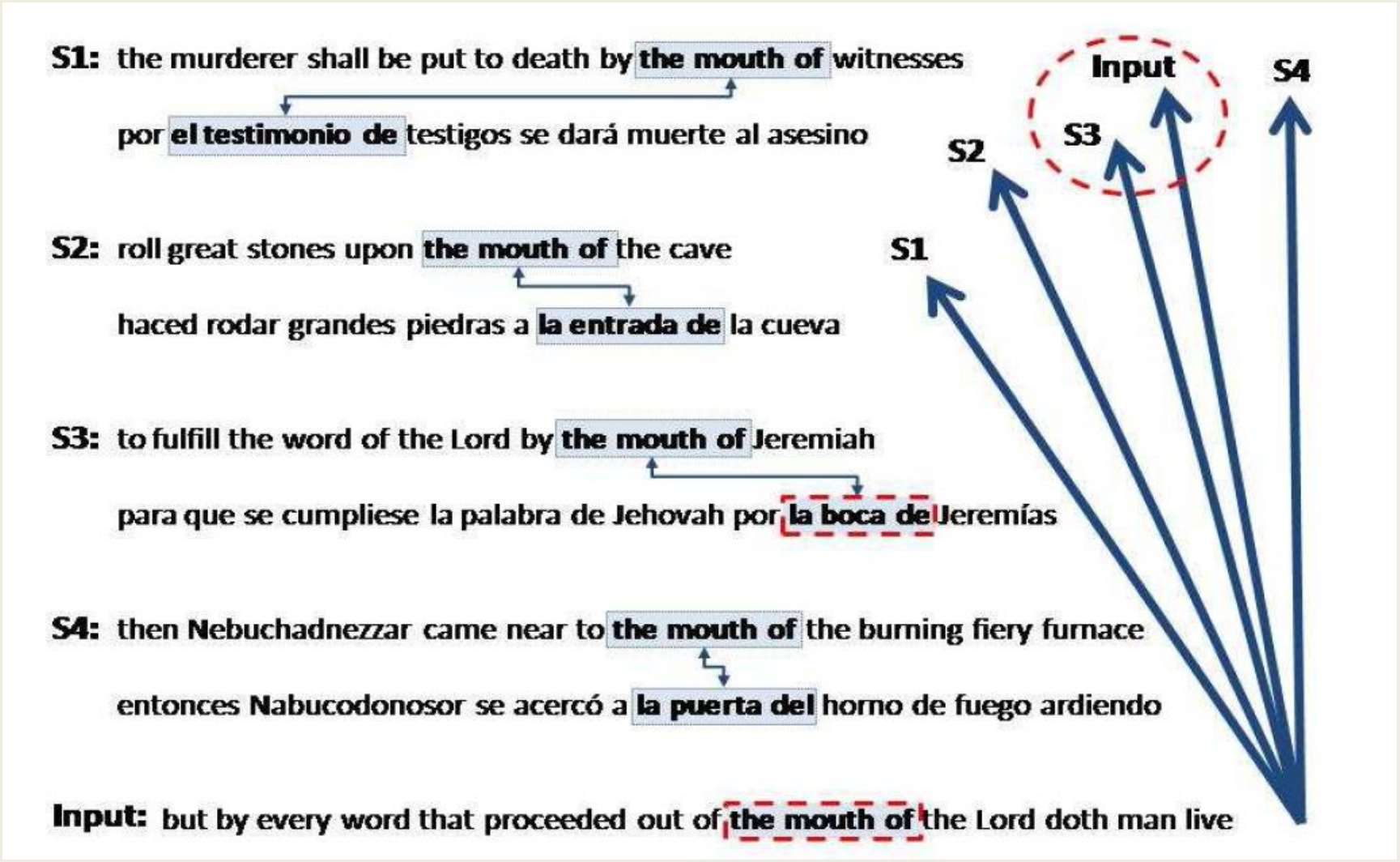
Banchs et al., 2011

Source-context Feature

- Let us consider a dynamic feature which value depends on the source-context of the input sentence to be translated.

$$F(TU,IN)=SIM(TU,IN)=SIM(SC,IN)$$

Illustrative Example of the Method

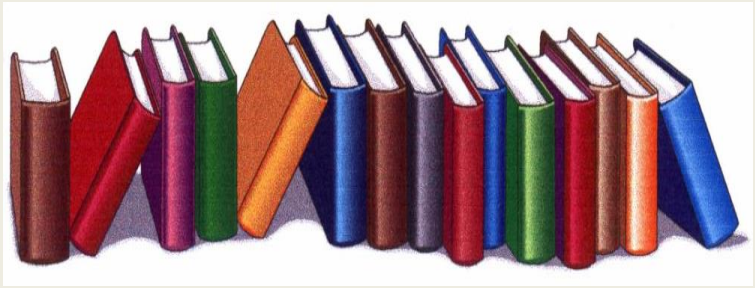


Sentence Similarity measures

- Vector space model
- Latent Semantic Analysis

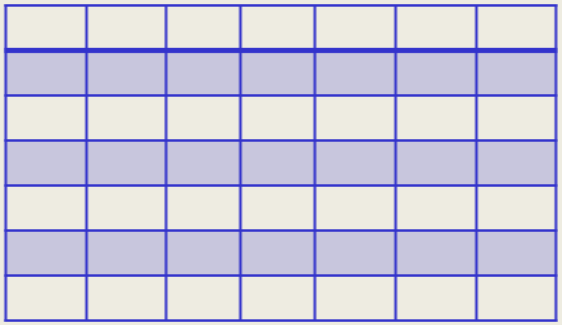
Term-document matrix

document collection



TS

Term-document matrix



MORE ADVANCED TOPICS: TF-IDF which is a numerical statistic which reflects how important a word is to a document in a collection or corpus.

Term-document example

Technical Memo Example

Titles:

c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

Term-document example

Terms	documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

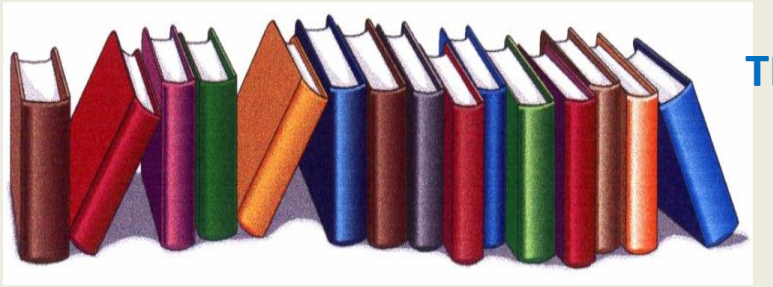
Further techniques: Latent Semantic Indexing

- LSI, the term-document matrix is decomposed into a set of K orthogonal factors by means of Singular Value Decomposition (SVD)

A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts

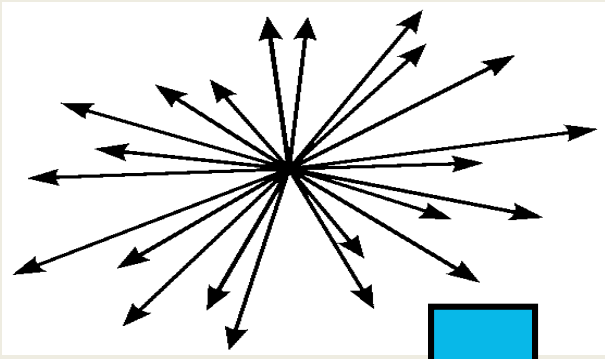
LSI applied to text data

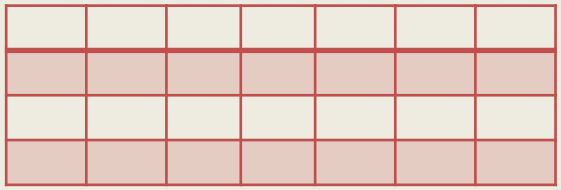
Document collection



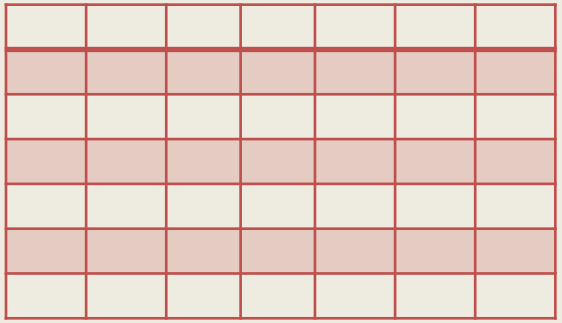
TF-IDF

Vector-space representation





LSI



Low-rank matrix

Tf-idf matrix

Computing sentence similarity

- A **vector space model** or **latent semantic analysis** can be used for computing distances or similarities between the input sentence to be translated and the original sentences phrases were extracted from.

Add the feature function in the phrase-based translation table

- Add a feature function that benefits translation units DERIVED from the TRAINING SENTENCES that are MORE SIMILAR to the SENTENCE TO TRANSLATE

		P_sim	P_mle	P_lex	.....
the mouth of	la boca de	0.98	0.54	0.32	
the mouth of	la puerta de	0.67	0.55	0.43	
the mouth of	la entrada de	0.57	0.53	0.34	
the mouth of	el testimonio de	0.45	0.52	0.23	

Adding context information using classification techniques

- Incorporate a local discriminative phrase selection model to address the semantic ambiguity
- Local classifiers are trained using linguistic and strutured context information to translate a phrase
- This significantly increases the accuracy in phrase selection

Discriminative Phrase Selection

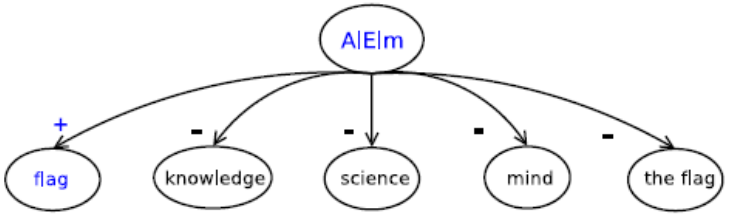
- Phrase selection is treated as a classification problem
- Use a classifier to solve the multiclass classification problem
- Training set: phrase-aligned parallel corpus
- Set of features: bag-of-words, local collocations, position-sensitive POS tags, basic dependency features…

Carpuat et al., 2008

Fragment of the translation table to take into account DPT predictions.

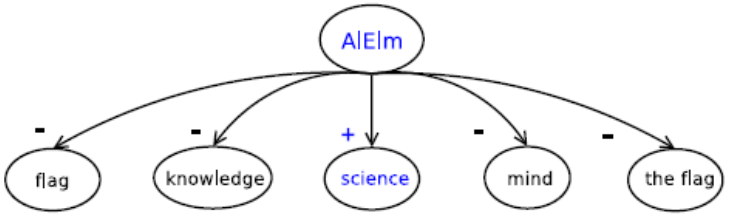
f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(f\|e)$	$lex(f\|e)$	$P_{MLE}(e\|f)$	$lex(e\|f)$
AlElm_1	flag	0.1986	0.6438	0.5417	0.3241	0.2826
AlElm_1	the	0.0419	0.0001	0.0001	0.0207	0.0217
AlElm_1	mind	0.0401	0.0608	0.0425	0.0620	0.0543
AlElm_1	the flag	0.0397	0.4000	0.5417	0.0414	0.0786
AlElm_1	flag during	0.0394	0.6667	0.5417	0.0138	0.0001
AlElm_1	knowledge	0.0392	0.0846	0.0798	0.1103	0.0924
AlElm_1	flag caused	0.0387	1.0000	0.5417	0.0138	0.0001
AlElm_1	science	0.0377	0.1529	0.1477	0.1793	0.1413
AlElm_1	education	0.0377	0.0018	0.0029	0.0138	0.0163
AlElm_1	in mind	0.0371	0.0571	0.0425	0.0138	0.0004
AlElm_1	...					

wAn$d AllbnAnywn Al*yn HmlwA ktb SlAp w rfEwA AlElm AllbnAny, Aln$yd AlwTny AllbnAny.



The Lebanese, who came carrying prayer books and the Lebanese flag, sang the Lebanese national anthem.

>n HAlp AlElm w Altknwlwjya ldY nA fy nhAyp Alqrn AlE$ryn l hA ElAmtAn mhmtAn. Al>wly gyAb AlmlAHqp fy h*A AlqTAE.



The situation of science and technology in Egypt at the end of the 20th century had two important features.

Binary classification with SVM

Every possible translation is a class ! one-vs-all Classification

España et al., 2010

Set of features: use context and linguistic information

- Features set for the SVMs include:
  - Source phrase features
    - PoS, coarse PoS and chunk n-grams
  - Source sentence features
    - Word, PoS, coarse PoS, chunk n-grams and bag-of-words

Set of features to train the SVM classifier

Annotated sentence ($\text{word}_{PoS|coarsePoS|chunk}$):

$\text{w}_{CC|C|O}$ $\text{tAbE}_{VBD|V|B-VP}$ $\text{mr\$d}_{NN|N|B-NP}$ $\text{AlIxwAn}_{NN|N|B-NP}$ $\text{``}_{PUNC|P|O}$ $\text{In}_{IN|I|B-SBAR}$ $\text{AlElm}_{NN|N|B-NP}$ $\text{AlmTlwb}_{JJ|J|I-NP}$ $\text{fy}_{IN|I|B-PP}$ $\text{dyn}_{NN|N|B-NP}$ $\text{nA}_{PRP\$|P|I-NP}$ $\text{hw}_{PRP|P|B-NP}$ $\text{kl}_{NN|N|B-NP}$ $\text{Elm}_{NN|N|I-NP}$ $\text{nAfE}_{NN|N|B-NP}$ ...

Phrase features:

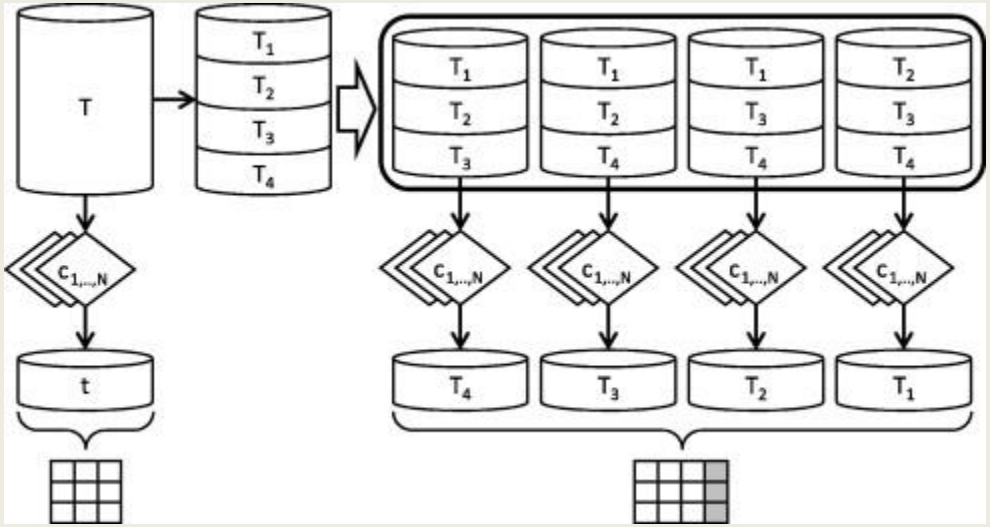
PoS	NN
coarse PoS	N
chunk	B-NP

Sentence features:

word	$(\text{AlmTlwb})_1$, $(\text{fy})_2$, $(\text{dyn})_3$, $(\text{nA})_4$, $(\text{hw})_5$,
n-grams	$(\text{In})_{-1}$, $(\text{"})_{-2}$, $(\text{AlIxwAn})_{-3}$, $(\text{mr\$d})_{-4}$, $(\text{tAbE})_{-5}$,
	$(\text{AlmTlwb fy})_1$, $(\text{fy dyn})_2$, $(\text{dyn nA})_3$, $(\text{nA hw})_4$,
	$(\text{In AlmTlwb})_{-1}$, $(\text{" In})_{-2}$, $(\text{AlIxwAn "})_{-3}$, $(\text{mr\$d AlIxwAn})_{-4}$,
	$(\text{tAbE mr\$d})_{-5}$,
	$(\text{AlmTlwb fy dyn})_1$, $(\text{fy dyn nA})_2$, $(\text{dyn nA hw})_3$,
	$(\text{In AlmTlwb fy})_{-1}$, $(\text{" In AlmTlwb})_{-2}$, $(\text{AlIxwAn " In})_{-3}$,
	$(\text{mr\$d AlIxwAn "})_{-4}$, $(\text{tAbE mr\$d AlIxwAn})_{-5}$
PoS	$(\text{JJ})_1$, $(\text{IN})_2$, $(\text{NN})_3$, $(\text{PRP\$})_4$, $(\text{PRP})_5$,
n-grams	$(\text{IN})_{-1}$, $(\text{PUNC})_{-2}$, $(\text{NN})_{-3}$, $(\text{NN})_{-4}$, $(\text{VBD})_{-5}$
	$(\text{JJ IN})_1$, $(\text{IN NN})_2$, $(\text{NN PRP\$})_3$, $(\text{PRP\$ PRP})_4$,
	$(\text{IN JJ})_{-1}$, $(\text{PUNC IN})_{-2}$, $(\text{NN PUNC})_{-3}$, $(\text{NN NN})_{-4}$, $(\text{VBD NN})_{-5}$
	$(\text{JJ IN NN})_1$, $(\text{IN NN PRP\$})_2$, $(\text{NN PRP\$ PRP})_3$,
	$(\text{IN JJ IN})_{-1}$, $(\text{PUNC IN JJ})_{-2}$,
	$(\text{NN PUNC IN})_{-3}$, $(\text{NN NN PUNC})_{-4}$, $(\text{VBD NN NN})_{-5}$,
coarse PoS	$(\text{J})_1$, $(\text{I})_2$, $(\text{N})_3$, $(\text{P})_4$, $(\text{P})_5$, $(\text{I})_{-1}$, $(\text{P})_{-2}$, $(\text{N})_{-3}$, $(\text{N})_{-4}$, $(\text{V})_{-5}$
n-grams	$(\text{J I})_1$, $(\text{I N})_2$, $(\text{N P})_3$, $(\text{P P})_4$,
	$(\text{I J})_{-1}$, $(\text{P I})_{-2}$, $(\text{N P})_{-3}$, $(\text{N N})_{-4}$, $(\text{V N})_{-5}$
	$(\text{J I N})_1$, $(\text{I N P})_2$, $(\text{N P P})_3$,
	$(\text{I J I})_{-1}$, $(\text{P I J})_{-2}$, $(\text{N P I})_{-3}$, $(\text{N N P})_{-4}$, $(\text{V N N})_{-5}$
chunk	$(\text{I-NP})_1$, $(\text{B-PP})_2$, $(\text{B-NP})_3$, $(\text{I-NP})_4$, $(\text{B-NP})_5$,
n-grams	$(\text{B-SBAR})_{-1}$, $(\text{O})_{-2}$, $(\text{B-NP})_{-3}$, $(\text{B-NP})_{-4}$, $(\text{B-VP })_{-5}$
	$(\text{I-NP B-PP})_1$, $(\text{B-PP B-NP})_2$, $(\text{B-NP I-NP})_3$, $(\text{I-NP B-NP})_4$,
	$(\text{B-SBAR I-NP})_{-1}$, $(\text{O, B-SBAR})_{-2}$, $(\text{B-NP O})_{-3}$, $(\text{B-NP B-NP })_{-4}$,
	$(\text{B-VP B-NP})_{-5}$
	$(\text{I-NP B-PP B-NP})_1$, $(\text{B-PP B-NP I-NP})_2$, $(\text{B-NP I-NP B-NP})_3$,
	$(\text{B-SBAR I-NP B-PP})_{-1}$, $(\text{O B-SBAR I-NP})_{-2}$, $(\text{B-NP O B-SBAR})_{-3}$,
	$(\text{B-NP B-NP O})_{-4}$, $(\text{B-VP B-NP B-NP })_{-5}$
bag-of-words	left: AlIxwAn, mr$d, tAbE
	right: $rEyAF, AlmTlwb, AlnAs, Elm, ElmAF, dyn, kAn, kl, nAfE, swA', tbgY, tjrybyAF, vmrt

Estimation of the discriminative phrase translation model and integration into the SMT system

- Training linear SVMs for every translation of every phrase
- Convert SVM score into probability via a softmax function
- Include this probability in the translation model within a Log-linear model



Memory-based multiclass classifier

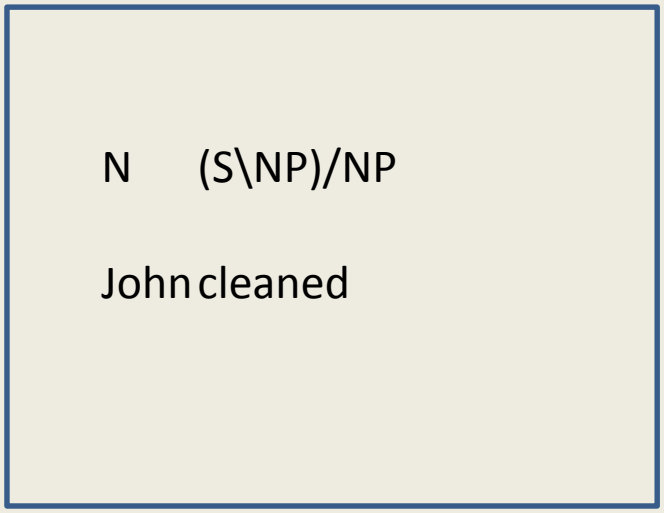
Features are used to train a multiclass classifier: Memory-based Learner TiMBL

Rejwanul, 2011

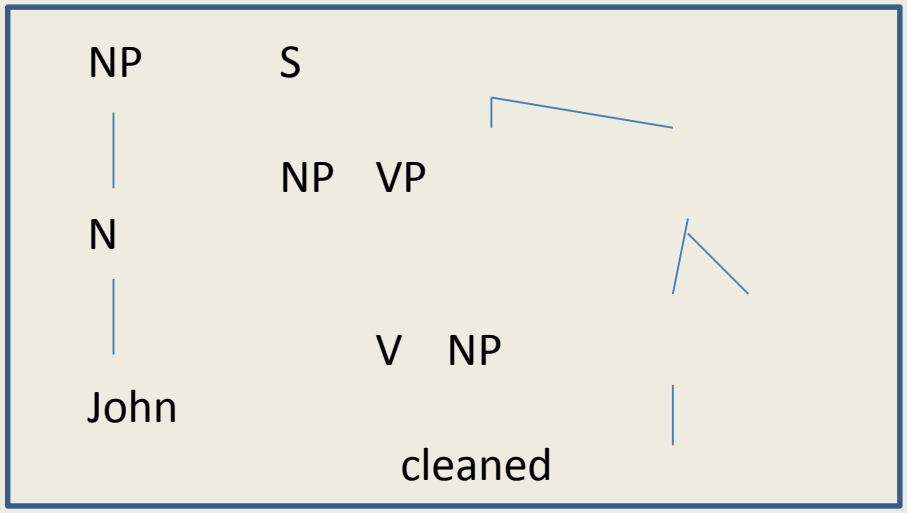
Set of features: structured context Information

- **Supertags**
  - **Combinatory Categorical Grammar (CCG)**
  - **Lexicalized Tree Adjoining Grammar (LTAG)**

CCG



LTAG



SHALLOW SEMANTICS

Using relationships between words

- Consider ***semantic role labeling*** (also known as shallow semantic parsing)

  [A0 He] [AM-MOD would] [AM-NEG n't]
  [V **accept**] [A1 anything of value] from
  [A2 those he was writing about] .

- Goal:
  - Identify predicates (e.g. *accept*)
  - Identify all the semantic rules such as subject, object, and indirect object
- Captures how words relate to one another

Importance of relationships

- Source: Anna felht ihrem Kater
- Output: Anna is missing her cat
- Reference: Anna's cat is missing her

- Words almost identical, but meaning is not!
  - German uses morphology to express roles
  - English uses word order
  - Need to capture and transfer this information

Integration of syntax and semantics

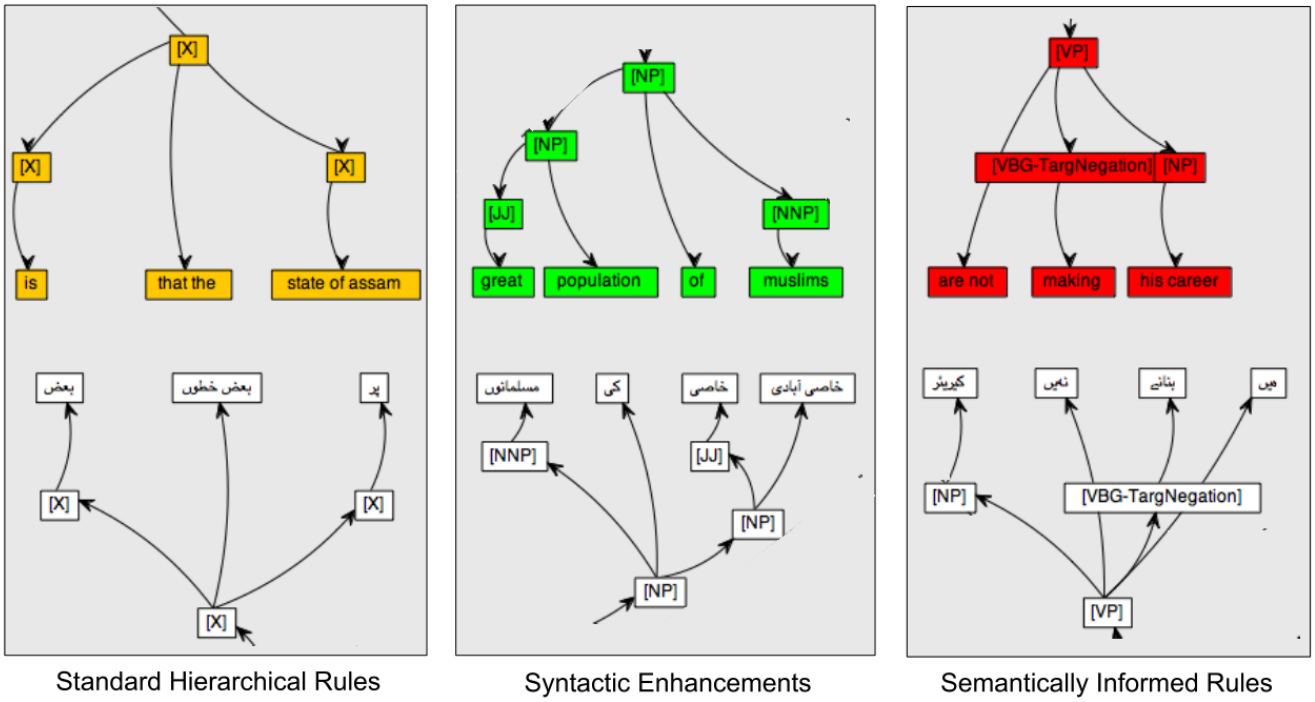


Figure 1.3: The evolution of HIVE integration in the Joshua decoder. At the start of summer the decoder used translation rules with a single generic non-terminal symbol, later syntactic categories were used, and by the end of the summer the translation rules included semantic entities such as modalities.

Syntax tree and semantic roles…

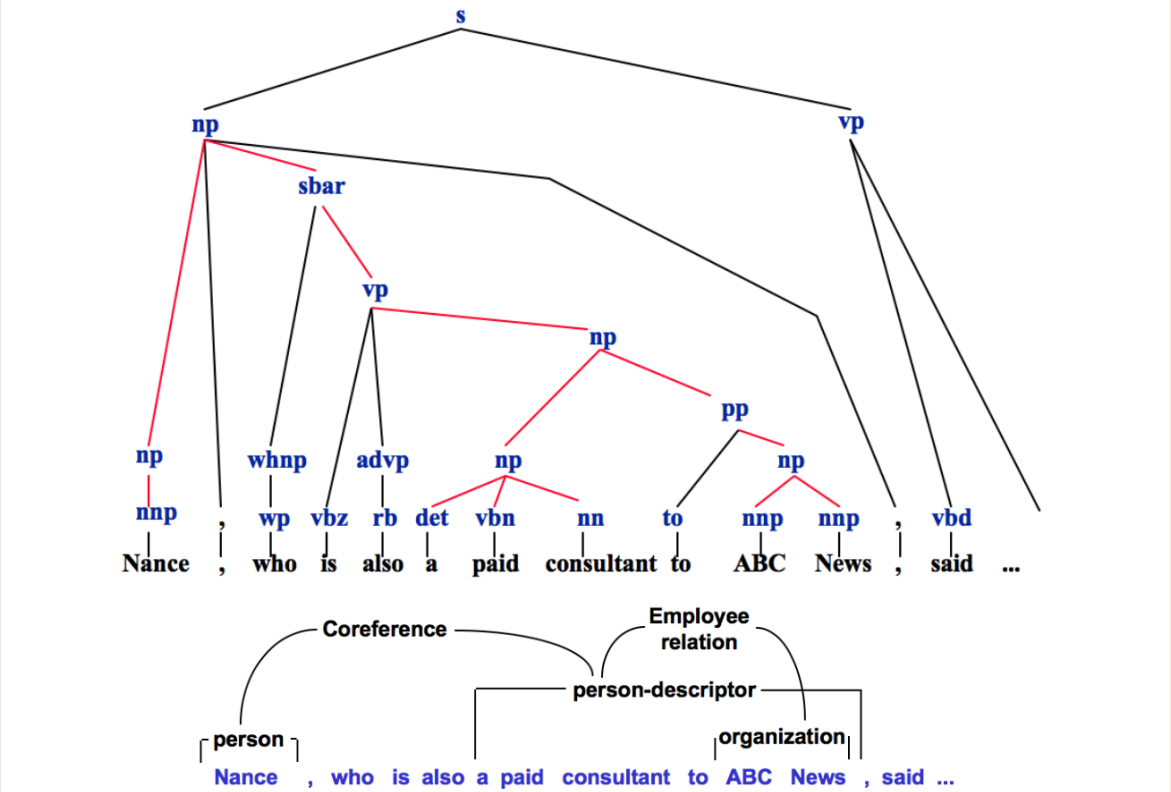


Figure 2.8: A sentence on the English side of the bilingual parallel training corpus is parsed with a syntactic parser, and tagged with a name finder.

…can be assembled together

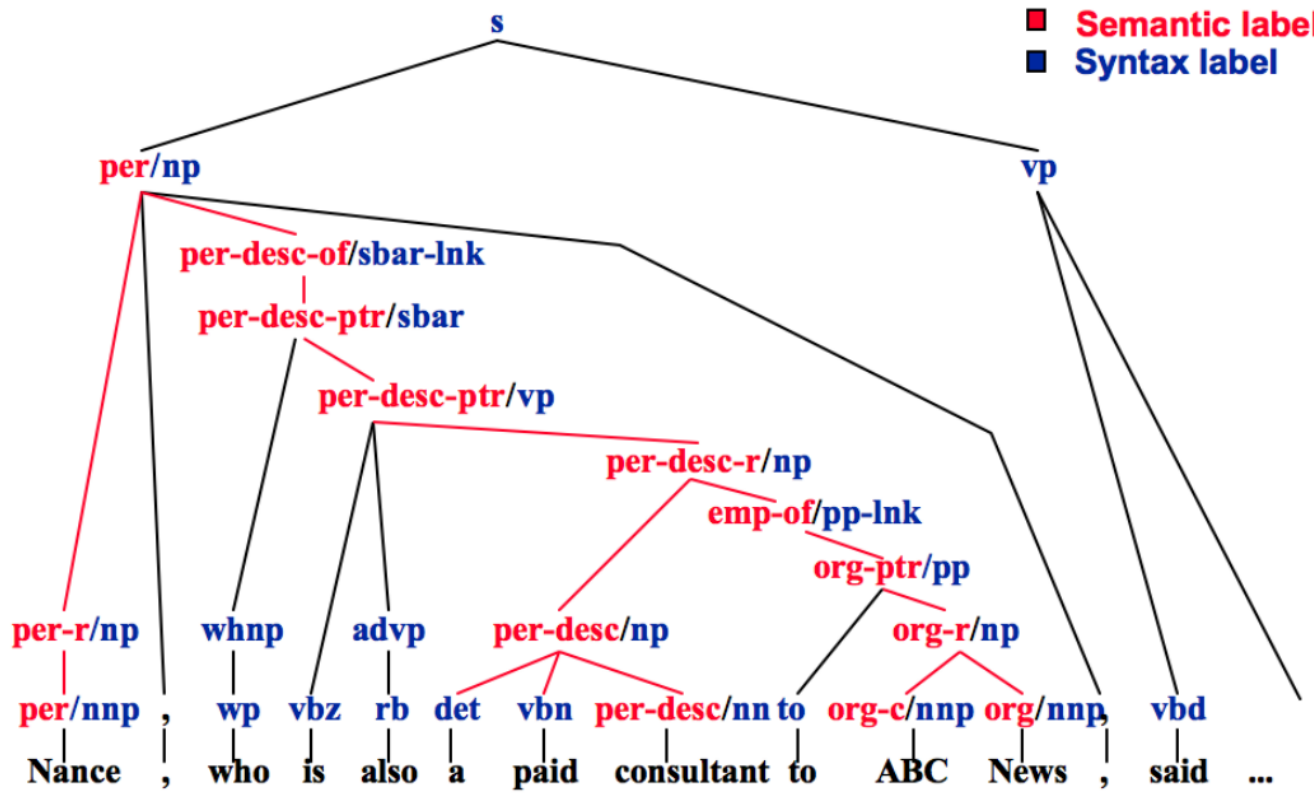


Figure 2.9: The name tags are grafted onto the syntactic parse tree prior to grammar extraction.

Some experimental results

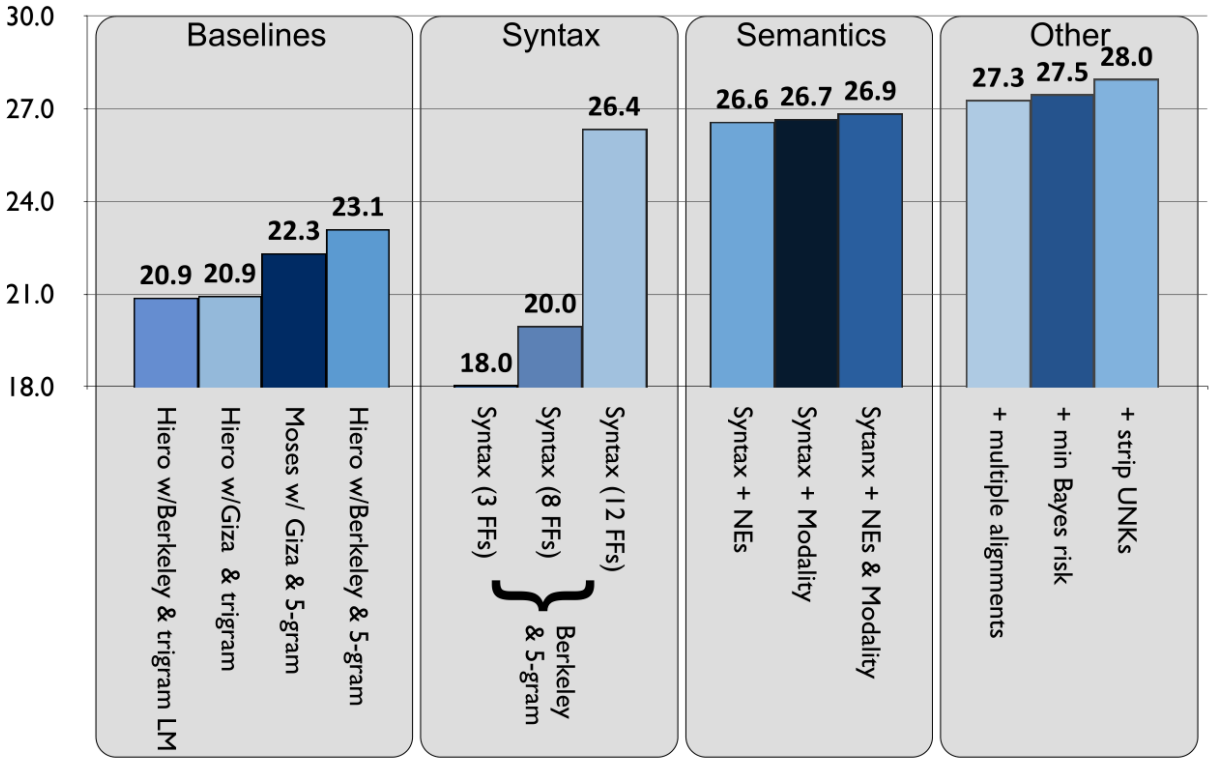
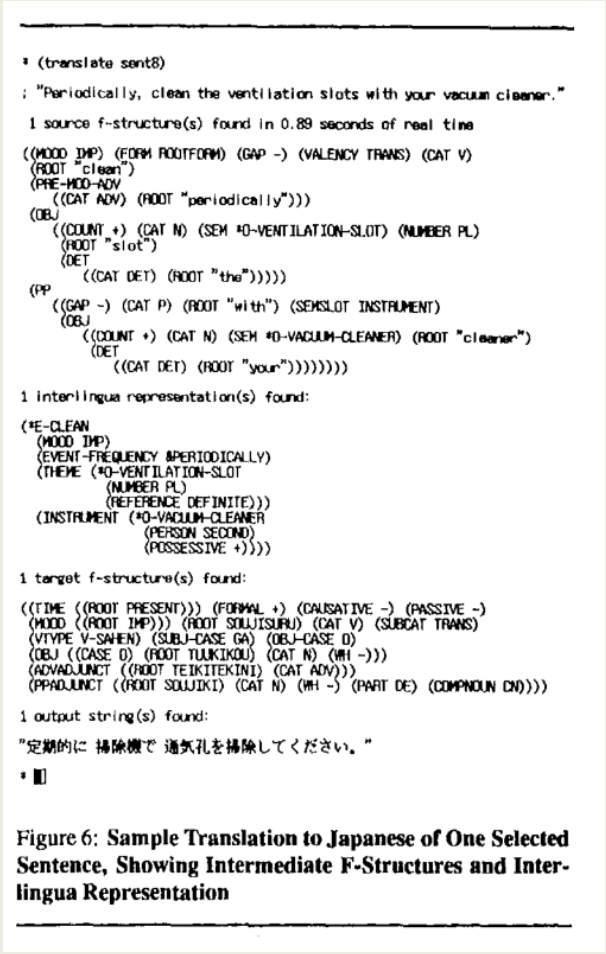


Figure 2.15: Results for a range of experiments conducted using Joshua during the SCALE workshop. The scores are lowercased Bleu calculated on the held-out devtest set.

FULL SENTENCE SEMANTICS

Classical interlingua approach: KANT: CMU 1980s-1990s



```
* (translate sent8)

; "Periodically, clean the ventilation slots with your vacuum cleaner."

 1 source f-structure(s) found in 0.89 seconds of real time

((MOOD IMP) (FORM ROOTFORM) (GAP -) (VALENCY TRANS) (CAT V)
 (ROOT "clean")
 (PRE-MOD-ADV
    ((CAT ADV) (ROOT "periodically")))
 (OBJ
    ((COUNT +) (CAT N) (SEM *O-VENTILATION-SLOT) (NUMBER PL)
     (ROOT "slot")
     (DET
        ((CAT DET) (ROOT "the")))))
 (PP
    ((GAP -) (CAT P) (ROOT "with") (SEMSLOT INSTRUMENT)
     (OBJ
        ((COUNT +) (CAT N) (SEM *O-VACUUM-CLEANER) (ROOT "cleaner")
         (DET
            ((CAT DET) (ROOT "your")))))))))

1 interlingua representation(s) found:

(*E-CLEAN
  (MOOD IMP)
  (EVENT-FREQUENCY &PERIODICALLY)
  (THEME (*O-VENTILATION-SLOT
           (NUMBER PL)
           (REFERENCE DEFINITE)))
  (INSTRUMENT (*O-VACUUM-CLEANER
                (PERSON SECOND)
                (POSSESSIVE +))))

1 target f-structure(s) found:

((TIME ((ROOT PRESENT))) (FORMAL +) (CAUSATIVE -) (PASSIVE -)
 (MOOD ((ROOT IMP))) (ROOT SOUJISURU) (CAT V) (SUBCAT TRANS)
 (VTYPE V-SAHEN) (SUBJ-CASE GA) (OBJ-CASE O)
 (OBJ ((CASE O) (ROOT TUUKIKOU) (CAT N) (WH -)))
 (ADVADJUNCT ((ROOT TEIKITEKINI) (CAT ADV)))
 (PPADJUNCT ((ROOT SOUJIKI) (CAT N) (WH -) (PART DE) (COMPNOUN CN))))

1 output string(s) found:

"定期的に 掃除機で 通気孔を掃除してください。"

*
```

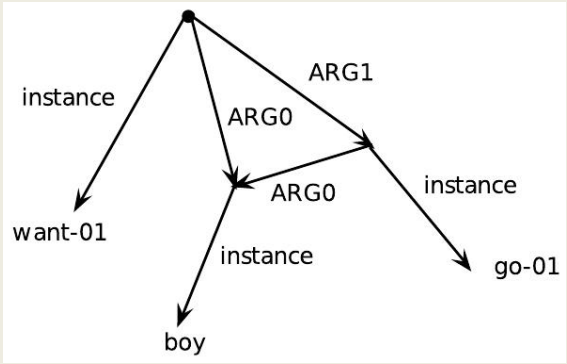
Figure 6: **Sample Translation to Japanese of One Selected Sentence, Showing Intermediate F-Structures and Interlingua Representation**

- Compelling aspects:
  - One parser and generator per language, instead of n^2 translation systems
- Problems:
  - Representation is very difficult! Languages capture different information
  - Mappings are very error prone

Recent resurgence of interest in semantics

- http://amr.isi.edu/language.html
  - As a simple example, they represent "The boy wants to go" as:

(w / want-01
  :ARG0 (b / boy)
  :ARG1 (g / go-01
    :ARG0 b))



- Building statistical models to
  - Rewrite semantics as strings (translation models)
  - Gauge likelihood of semantics (language model)

Semantic approaches in statistical machine translation

CONCLUSIONS

Semantic approaches

- Benefits and integration of :
  - WSD methods
  - IR methods

into a statistical machine translation system

Main idea: lexical semantics

Translation unit from:

- P(e|f)
- SOURCE:TARGET

into

- P(e|f, CI(f))
- SOURCE : TARGET: SOURCE CONTEXT

Additional knowledge and algorithms

Knowledge

- Similarity measures
- Bag-of-words
- Part-of-speech
- Dependencies
  - Supertags

Algorithms

- Classifiers
  - SVM
  - Memory-based classifiers
- Vector-space model
- Latent semantic analysis

Any type of source language contexts are effective

Advantages vs Disadvantages

- Consistent improvements
- Different linguistic levels improvements
- Domain independent
- New tools appearing: suffix arrays + GPUs

- Difficult to integrate with PBSMT
- Computationallly expensive
- Not tested open-source tools available (starting…)

Semantic structures

- Early work has shown small gains in practice
- Many difficult problems to surmount
  - What are the best semantic representations?
  - How can we accurately map surface strings into semantic representations?

References

- BANCHS, R. E. AND COSTA-JUSSÀ, M. R. 2011. **A semantic feature for statistical machine translation**. In Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation. SSST-5. Association for Computational Linguistics, Stroudsburg, PA, USA, 126–134.
- CARPUAT, M. AND WU, D. 2008. **Evaluation of context-dependent phrasal translation lexicons for statistical machine translation**. In 6th International Conference on Language Resources and Evaluation (LREC-2008).
- ESPAÑA-BONET, C., GIMÉNEZ, J., AND MÀRQUEZ, L. 2010. **Discriminative phrase-based models for arabic machine translation**. ACM Transactions on Asian Language Information Processing Journal (TALIP) 8, 1–20.
- HAQUE, R. 2011. **Integrating source-language context into log-linear models of statistical machine translation.** Ph.D. thesis, Dublin City University