# Improving Students' Writing with Automated Grammatical Error Correction

Hwee Tou Ng

Department of Computer Science
National University of Singapore

15 October 2013

NUS National University of Singapore | School of Computing

Leading The World With Asia's Best

# Grammatical Error Correction (GEC)

- Task: Detect and correct grammatical errors
  - Input: English essays written by learners of English
  - Output: Corrected essays

# Sample Grammatical Errors

- Article or determiner
  - In late nineteenth century, …
  - late → the late
- Preposition
  - They must pay more on the welfare of the old people
  - on → for
- Noun number
  - Such powerful device shall not be made available.
  - device → devices

# Sample Grammatical Errors

- Verb form
  - Our society is progressed well.
  - progressed → progressing
- Subject-verb agreement
  - Some people still prefers to be single.
  - prefers → prefer

# Impact of GEC Research

# Impact of GEC Research

- More than one billion people worldwide are learning English as a second language
- More non-native English speakers than native speakers
- Of particular relevance in the Asian context
- A complete end-to-end application

# Historical Context

- Grammar checking is one of the first commercial NLP applications
- Microsoft Word Grammar Check
  - Heidorn, Jansen, et al. (IBM T J Watson, then Microsoft Research)
  - A hand-crafted rule-based approach
  - Limited coverage (detects none of the 5 sample grammatical errors shown)

# Current Landscape

- Commercial software available:

ETS Criterion

GINGER

Grammarly

WhiteSmoke
Just write.

Grammar Expert Plus

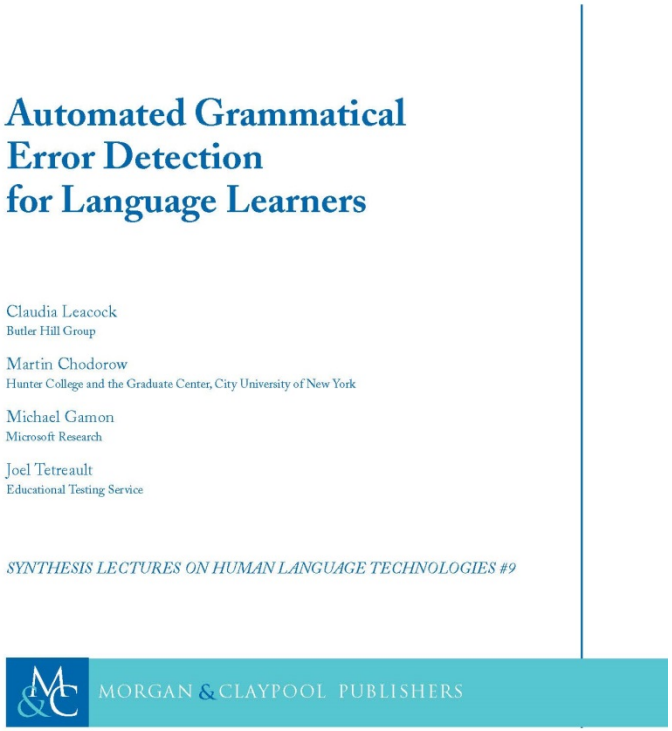
CorrectEnglish

# Current Landscape

- A somewhat neglected research topic
  - Relatively less published research in the NLP literature
- ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA) in 2003, 2005, 2008, 2009, 2010, 2011, 2012, 2013

# Introductory Book

Automated Grammatical Error Detection for Language Learners, Leacock, Chodorow, Gamon, Tetreault, 2010, Synthesis Lectures on Human Language Technologies



# State of the Art

- Up till 2010, unclear what that is
- Few annotated learner corpora for evaluation
- Existing corpora either small or proprietary

"… a reasonably sized public data set for evaluation and an accepted annotation standard are still sorely missing. Anyone developing such a resource and making it available to the research community would have a major impact on the field, …"

Leacock et al., 2010

# Shared Tasks on GEC

- Much recent research interest
- Three shared tasks:
  - Helping Our Own (HOO) 2011 (Dale and Kilgarriff, 2011)
  - Helping Our Own (HOO) 2012 (Dale et al., 2012)
  - CoNLL 2013 Shared Task (Ng et al., 2013)

# Automated Essay Scoring

- Task: output a single score only for an essay
- Different from grammatical error correction
- Less informative to a learner
- The Hewlett Foundation sponsored the Automated Student Assessment Prize (ASAP) in Feb – Apr 2012
- Recent work of Yannakoudakis, Briscoe, Medlock, ACL 2011

# HOO (Helping Our Own) 2011

- The first shared task on grammatical error correction
- Goal: Help NLP authors in writing their papers ("helping our own")
- Annotated corpus (publicly available):
  - Parts of 19 papers from the ACL Anthology
  - # of word tokens in development data = 22,806
  - # word tokens in test data = 18,789

# HOO 2011

- All error types (about 80) from the Cambridge University Press Error Coding System (Nicholls, 2003)
- Participants mostly address article and preposition errors only
- 6 participating teams
- Top performance: UIUC team (Rozovskaya, Sammons, Gioja, & Roth, 2011)

# HOO 2012

- Focus on determiner and preposition errors only
- Annotated corpus:
  - Cambridge FCE (First Certificate in English) exam scripts (part of the Cambridge Learner Corpus)
  - Training data (publicly available):
    - # scripts = 1,244
    - # words = 374,680
  - Test data (not available after the shared task):
    - # scripts = 100
    - # words = 18,013
- 14 participating teams
- Top performance: NUS team (D. Dahlmeier, H. T. Ng, & E. J. F. Ng, 2012)

# CoNLL-2013 Shared Task

- Input: English test essays
  - Pre-processed form provided (sentence segmentation, tokenization, POS tagging, constituency parsing, dependency parsing)
- Output: Corrected test essays, in sentence-segmented and tokenized form

# Training Data

- NUCLE corpus (NUS Corpus of Learner English) (Dahlmeier & Ng, 2011; Dahlmeier, Ng, & Wu, 2013)
- Publicly available for research purpose
  - http://www.comp.nus.edu.sg/~nlp/corpora.html
- Essays written by university students at NUS who are non-native speakers of English
- A wide range of topics (surveillance technology, health care, etc.)
- Hand-corrected by professional English instructors at NUS
- 27 error types

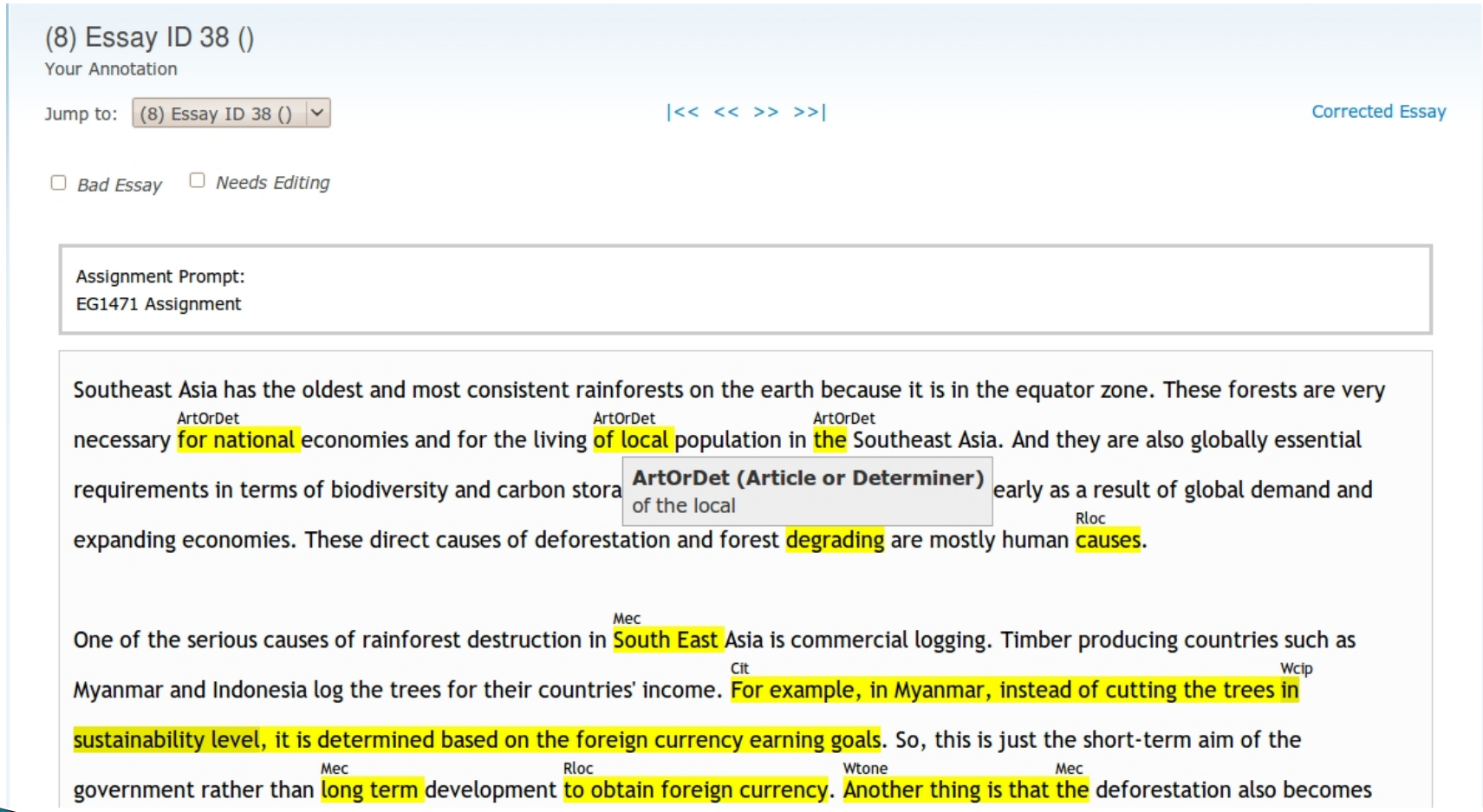
# NUCLE Error Types

| Error Tag | Error Type | Error Tag | Error Type |
|---|---|---|---|
| Vt | Verb tense | Srun | Runons, comma splice |
| Vm | Verb modal | Smod | Dangling modifier |
| V0 | Missing verb | Spar | Parallelism |
| Vform | Verb form | Sfrag | Fragment |
| SVA | Subject–verb agreement | Ssub | Subordinate clause |
| ArtOrDet | Article or determiner | WOinc | Incorrect sentence form |
| Nn | Noun number | WOadv | Adverb/adjective position |
| Npos | Noun possessive | Trans | Link words/phrases |
| Pform | Pronoun form | Mec | Punctuation, capitalization, spelling, typos |
| Pref | Pronoun reference | Rloc | Local redundancy |
| Wcip | Wrong collocation/idiom/preposition | Cit | Citation |
| Wa | Acronym | Others | Other errors |
| Wform | Word form | Um | Unclear meaning |
| Wtone | Tone | | |

# WAMP

- Writing, Annotation, and Marking Platform (WAMP)
- Online annotation tool developed at the NUS NLP group
- Used to create the NUCLE corpus

# WAMP


(8) Essay ID 38 ()
Your Annotation
Jump to: (8) Essay ID 38 ()
|<< << >> >>|
Corrected Essay
Bad Essay
Needs Editing
Assignment Prompt:
EG1471 Assignment
Southeast Asia has the oldest and most consistent rainforests on the earth because it is in the equator zone. These forests are very
ArtOrDet
necessary for national economies and for the living
ArtOrDet
of local population in
ArtOrDet
the Southeast Asia. And they are also globally essential
ArtOrDet (Article or Determiner)
of the local
requirements in terms of biodiversity and carbon stora early as a result of global demand and
expanding economies. These direct causes of deforestation and forest degrading are mostly human
Rloc
causes.
Mec
One of the serious causes of rainforest destruction in South East Asia is commercial logging. Timber producing countries such as
Cit
Wcip
Myanmar and Indonesia log the trees for their countries' income. For example, in Myanmar, instead of cutting the trees in
sustainability level, it is determined based on the foreign currency earning goals. So, this is just the short-term aim of the
Mec
Rloc
Wtone
Mec
government rather than long term development to obtain foreign currency. Another thing is that the deforestation also becomes

# A Sample Error Annotation

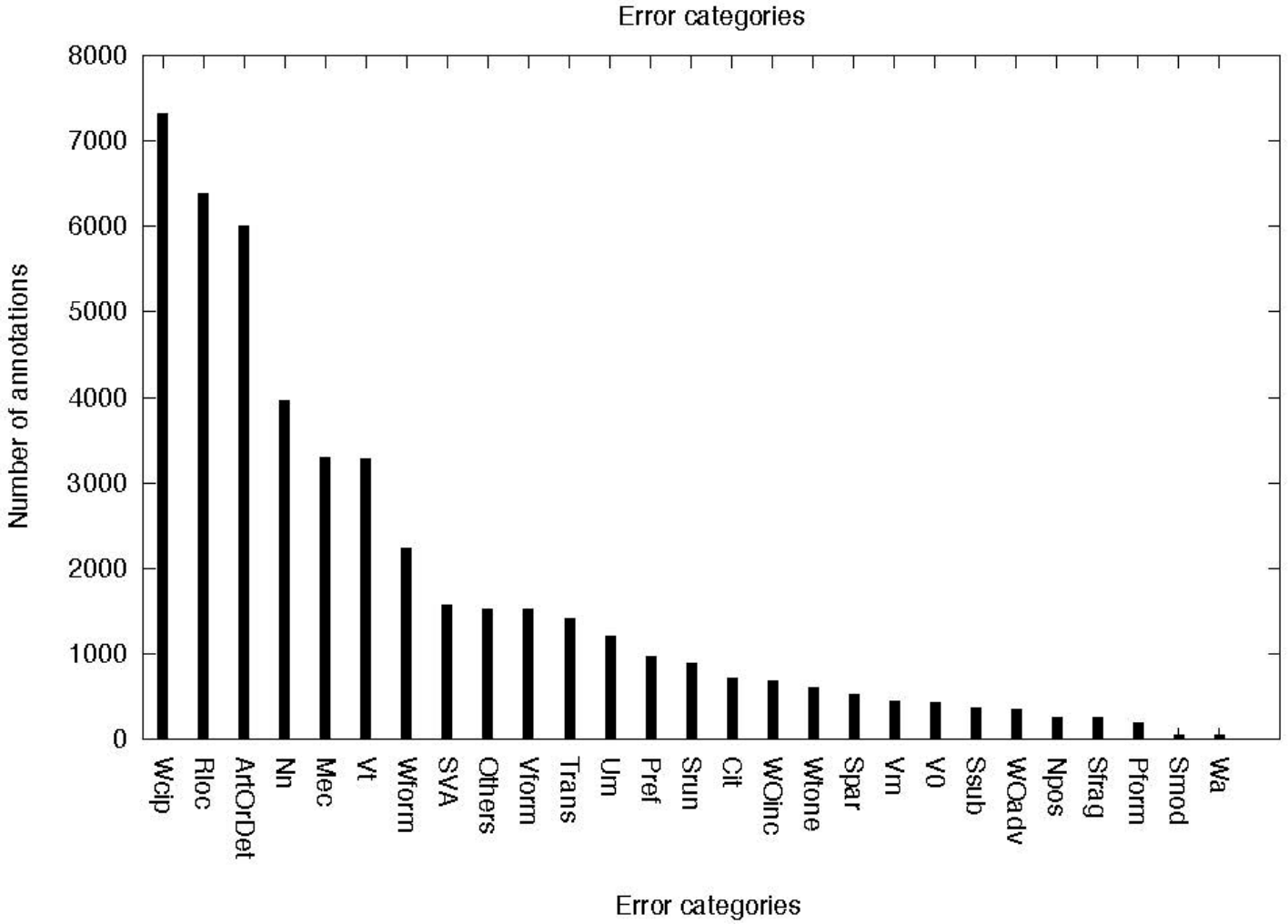
```
<MISTAKE start_par="0" start_off="5" end_par="0" end_off="9">
<TYPE>ArtOrDet</TYPE>
<CORRECTION>the past</CORRECTION>
</MISTAKE>
```

- Sentence:
  - From past to the present, …
  - past → the past
- Character offsets
- Stand-off annotations, in SGML format
- Error annotations automatically mapped to token offsets after pre-processing

# Statistics of NUCLE (version 2.3)

- # essays = 1,397
- # sentences = 57,151
- # word tokens = 1,161,567
- # errors (in all 27 error types) = 45,106

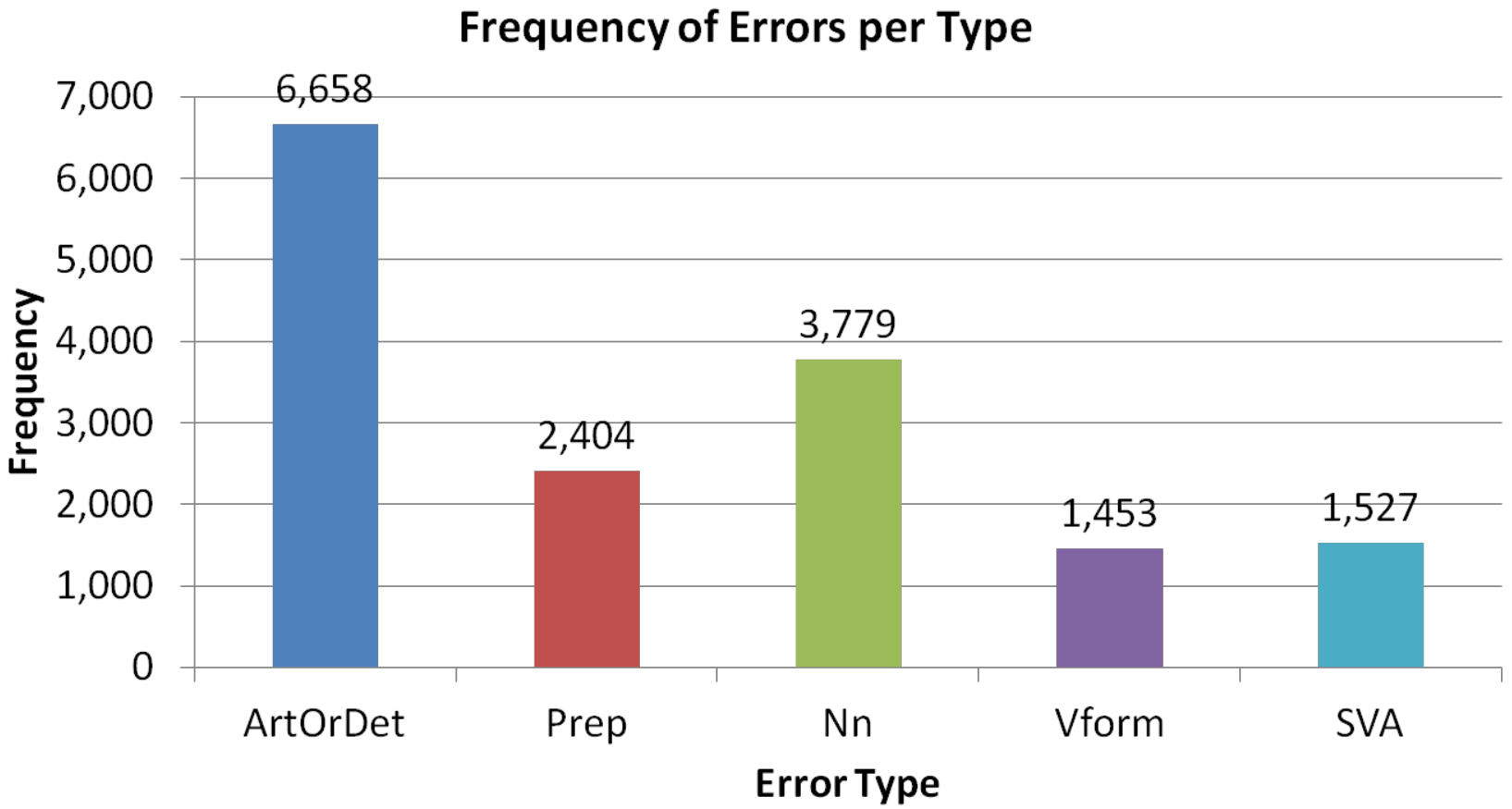
# Statistics of Errors in NUCLE



# CoNLL-2013 Task Definition

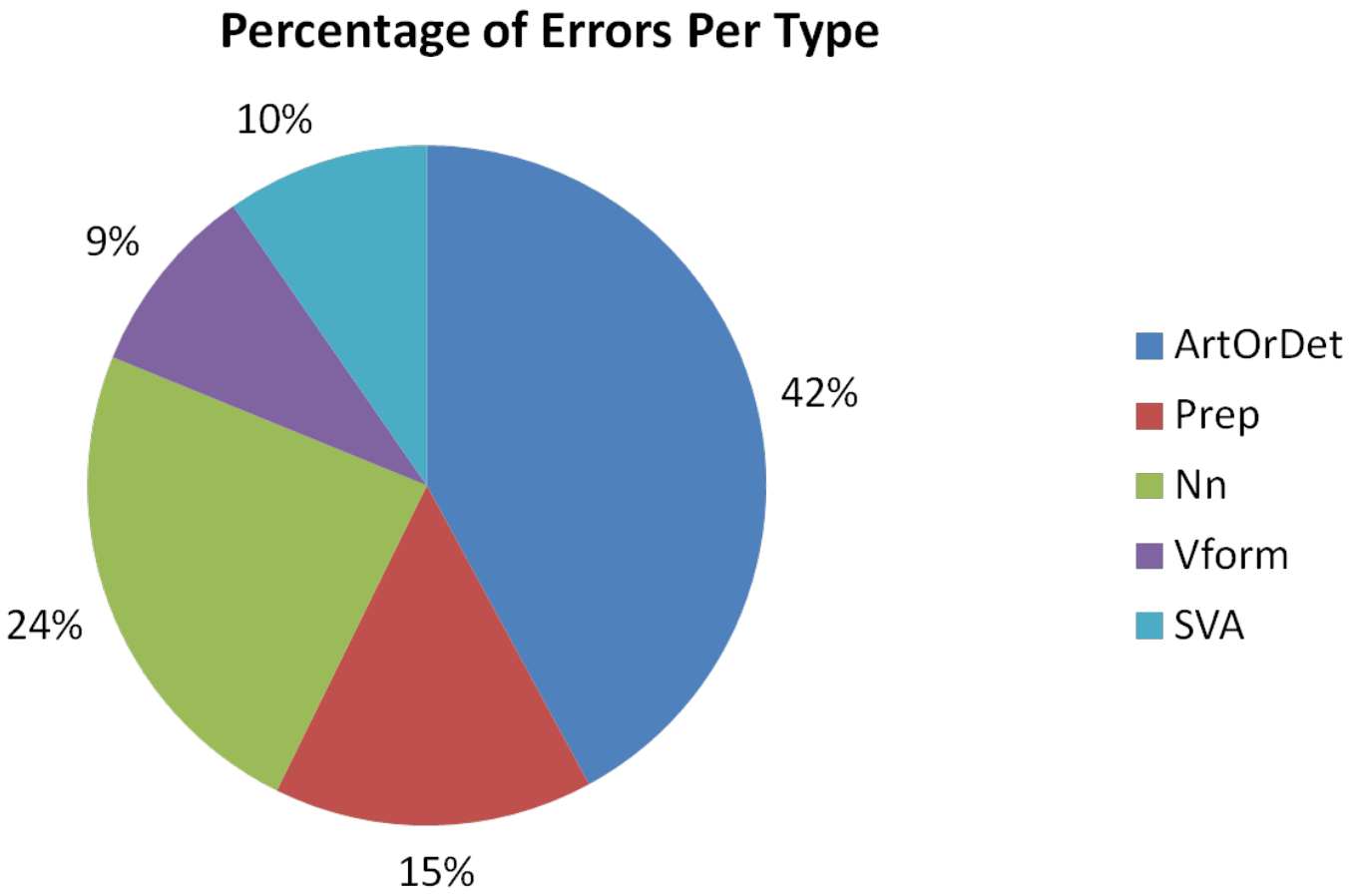
- Focus on 5 error types
  - Article or determiner (ArtOrDet)
  - Preposition (Prep)
  - Noun number (Nn)
  - Verb form (Vform)
  - Subject–verb agreement (SVA)
- Test essays still contain all errors, but corrections are made only on these 5 error types

# Statistics of Errors in NUCLE




Total number of errors of the 5 types = 15,821

# Statistics of Errors in NUCLE



# Test Data

- 50 new essays written by 25 NUS students (2 essays per student)
- Two prompts: one essay written for each prompt (one new prompt, one used in NUCLE)
- # sentences = 1,381
- # word tokens = 29,207

# Two Prompts for the Test Essays

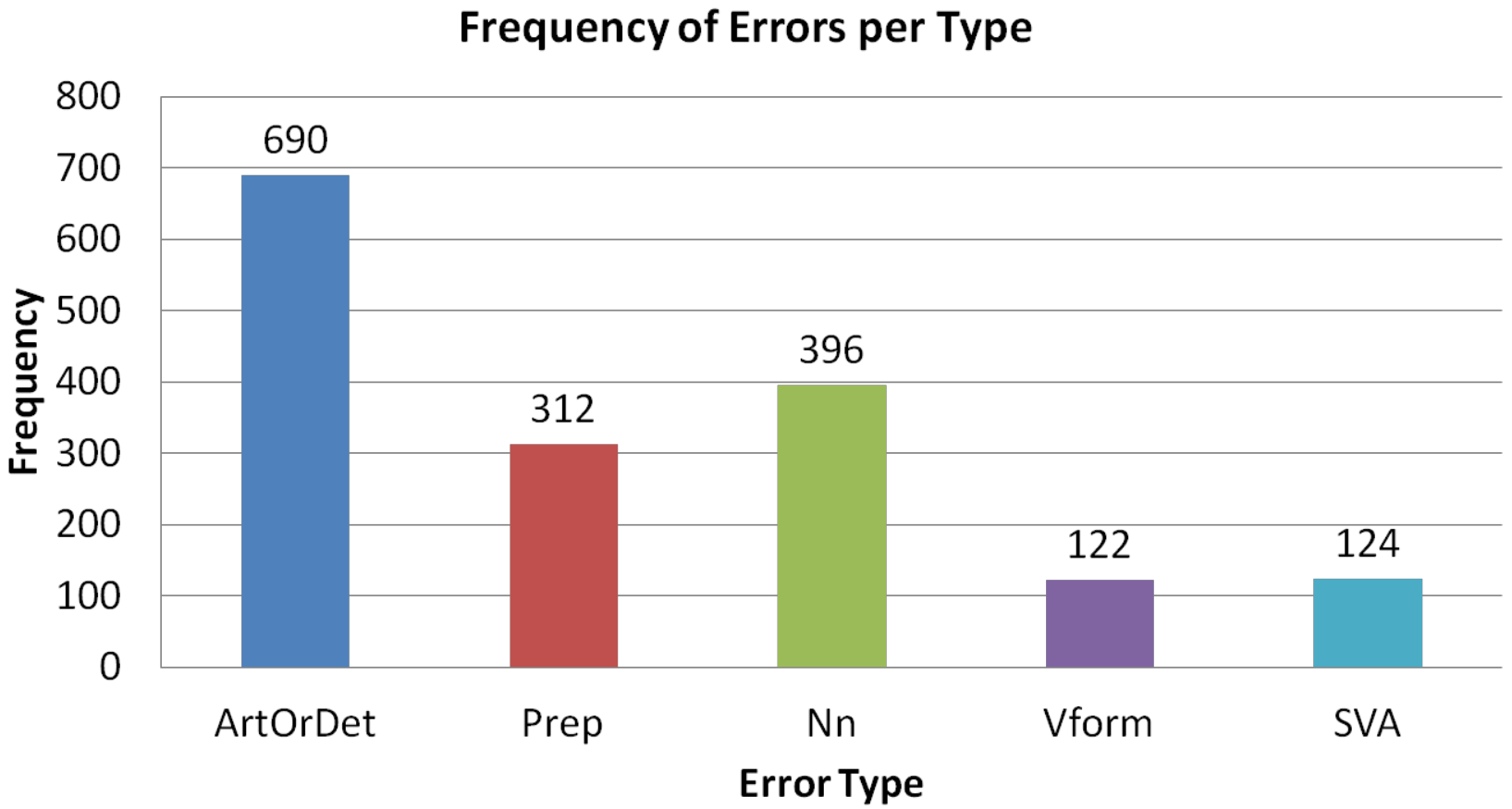
- Surveillance technology such as RFID (radio-frequency identification) should not be used to track people (e.g., human implants and RFID tags on people or products). Do you agree? Support your argument with concrete examples.
- Population aging is a global phenomenon. Studies have shown that the current average life span is over 65. Projections of the United Nations indicate that the population aged 60 or over in developed and developing countries is increasing at 2% to 3% annually. Explain why rising life expectancies can be considered both a challenge and an achievement.

# Test Data

- Annotation on test essays carried out by a native speaker of English (a lecturer at the NUS Centre for English Language Communication)
- Time spent on annotation: 25 hours
- Test essays and annotations freely available at the shared task home page:

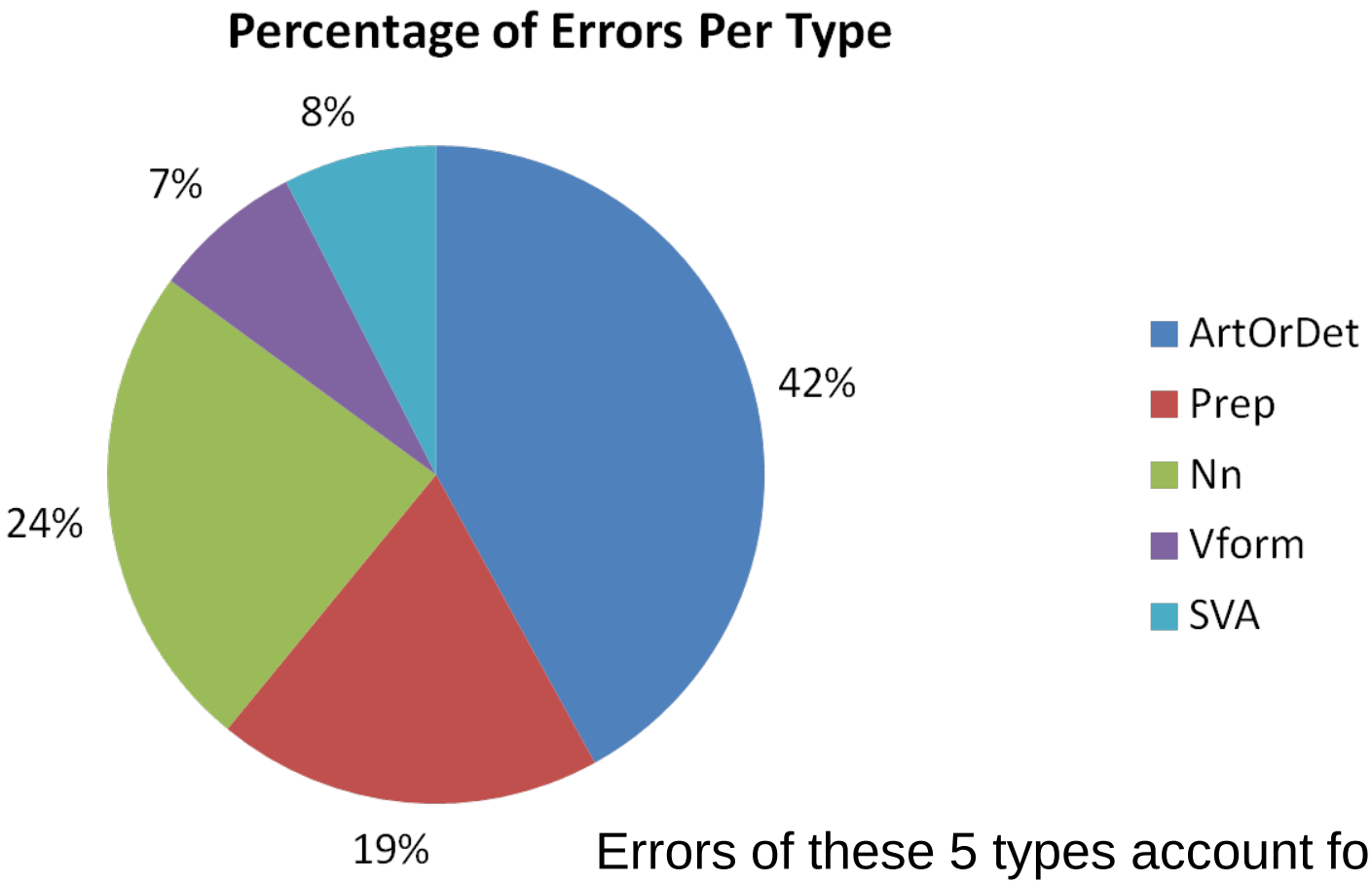
  http://www.comp.nus.edu.sg/~nlp/conll13st.html

# Statistics of Errors in Test Data




Total number of errors of the 5 types = 1,644

# Statistics of Errors in Test Data




Errors of these 5 types account for 47% of all errors in test essays

# Usage of Training Data and Tools

- Shared task participants are free to use other (or additional) corpora or tools, provided that they are publicly available

# Evaluation

- Edits: corrections
- How well the proposed system edits ($e_i$) match the gold-standard edits ($g_i$)
- Recall (R), Precision (P), F1 measure

$$R = \frac{\sum_{i=1}^{n} |\mathbf{g}_i \cap \mathbf{e}_i|}{\sum_{i=1}^{n} |\mathbf{g}_i|} \qquad P = \frac{\sum_{i=1}^{n} |\mathbf{g}_i \cap \mathbf{e}_i|}{\sum_{i=1}^{n} |\mathbf{e}_i|} \qquad F_1 = \frac{2 \times R \times P}{R + P}$$

# Evaluation

- Example:
  - Original sentence:
    - There is no a doubt , tracking system has brought many benefits .
  - Gold-standard edits g = { a doubt → doubt, system → systems, has → have }
  - Corrected sentence:
    - There is no doubt , tracking system has brought many benefits .
  - System edits e = { a doubt → doubt }
  - R = 1/3, P = 1/1, F1 = 1/2

# Anomaly of HOO Scorer

- Original sentence:
  - There is no a doubt , tracking system has brought many benefits .
- Gold-standard edits g = { a doubt → doubt, system → systems, has → have }
- Multiple, equivalent gold-standard edits
  - { a → ϵ, system → systems, has → have }
  - { a → ϵ, system has → systems have }
- Corrected sentence:
  - There is no doubt , tracking system has brought many benefits .
- GNU wdiff gives system edits e = { a → ϵ }
- HOO scorer gives erroneous scores: R = P = F1 = 0

# Scorer

- MaxMatch (M2) scorer (Dahlmeier & Ng, 2012)
- Automatically determine the system edits that maximally match the gold-standard edits
- Efficiently search for such system edits using an edit lattice
- Scorer can be freely downloaded from the shared task home page:

  http://www.comp.nus.edu.sg/~nlp/conll13st.html

# Participating Teams (17)

| Team ID | Affiliation |
|---|---|
| CAMB | University of Cambridge |
| HIT | Harbin Institute of Technology |
| IITB | Indian Institute of Technology, Bombay |
| KOR | Korea University |
| NARA | Nara Institute of Science and Technology |
| NTHU | National Tsing Hua University |
| SAAR | Saarland University |
| SJT1 | Shanghai Jiao Tong University (Team #1) |
| SJT2 | Shanghai Jiao Tong University (Team #2) |

| Team ID | Affiliation |
|---|---|
| STAN | Stanford University |
| STEL | Stellenbosch University |
| SZEG | University of Szeged |
| TILB | Tilburg University |
| TOR | University of Toronto |
| UAB | Universitat Autònoma de Barcelona |
| UIUC | University of Illinois at Urbana-Champaign |
| UMC | University of Macau |

Asia: 8
Europe/Africa: 6
North America: 3

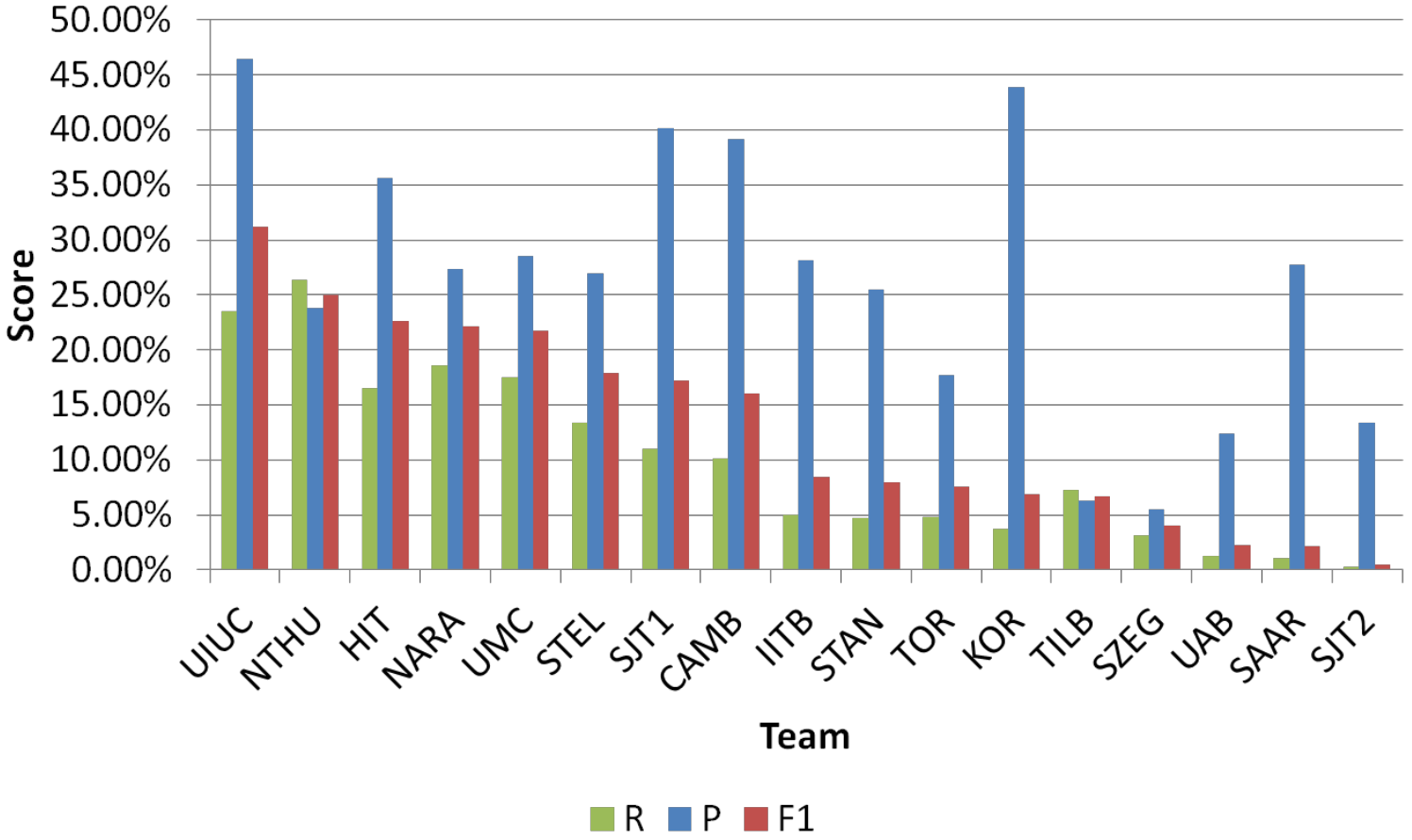
# Alternative Annotations

- Nature of grammatical error correction:
  - Multiple, different corrections are often acceptable
- Allow participants to raise their disagreement with the original gold-standard annotations
- Prevent under-estimation of performance
- Similar to HOO 2011 & HOO 2012
- Extend M2 scorer to deal with multiple alternative gold-standard annotations

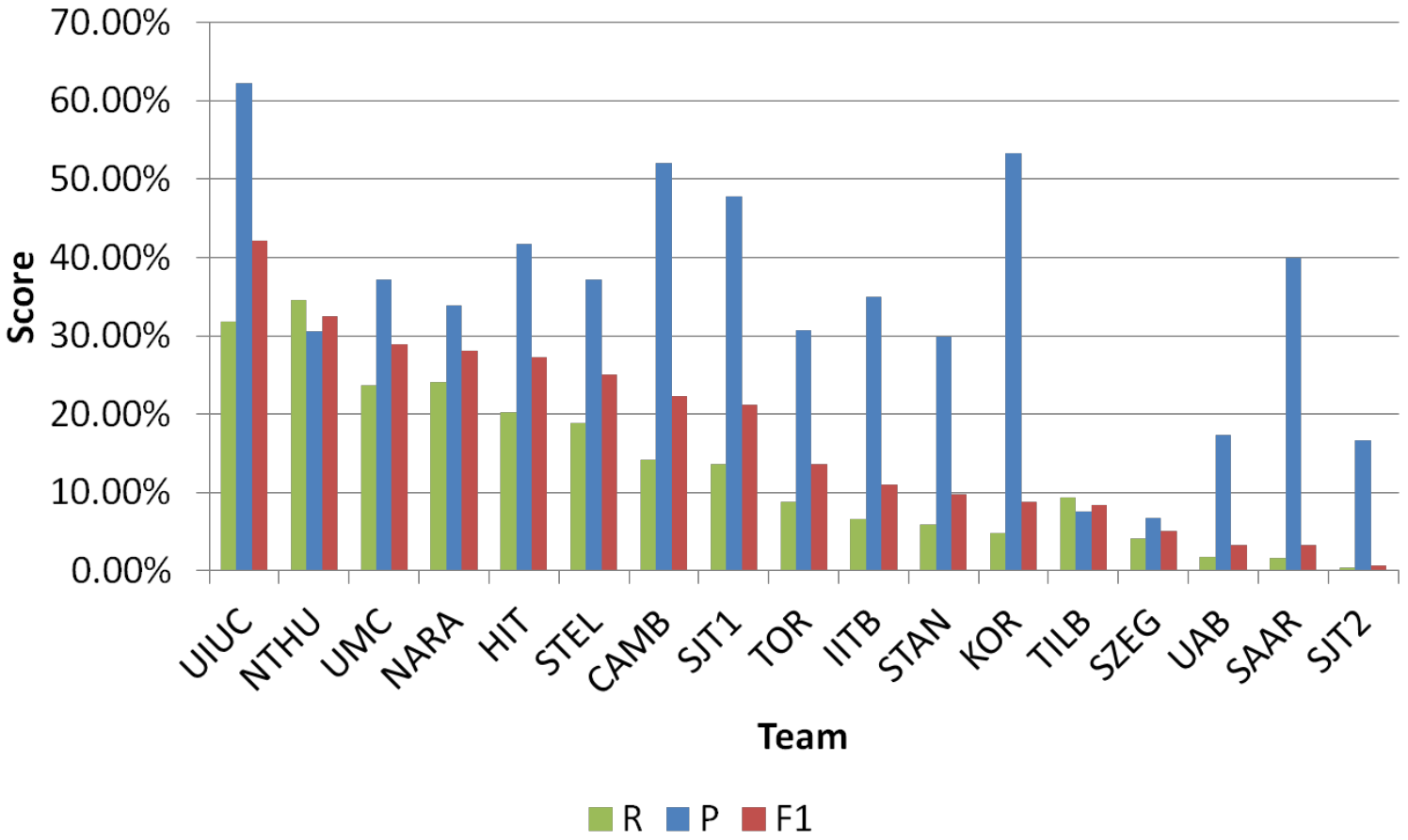
# Alternative Annotations

- Five teams (NTHU, STEL, TOR, UIUC, UMC) submitted alternative answers
- The same annotator who provided the original gold-standard annotations judged the alternative answers proposed (time spent: 17 hours)
- F1 scores of all teams improve when evaluated with alternative answers

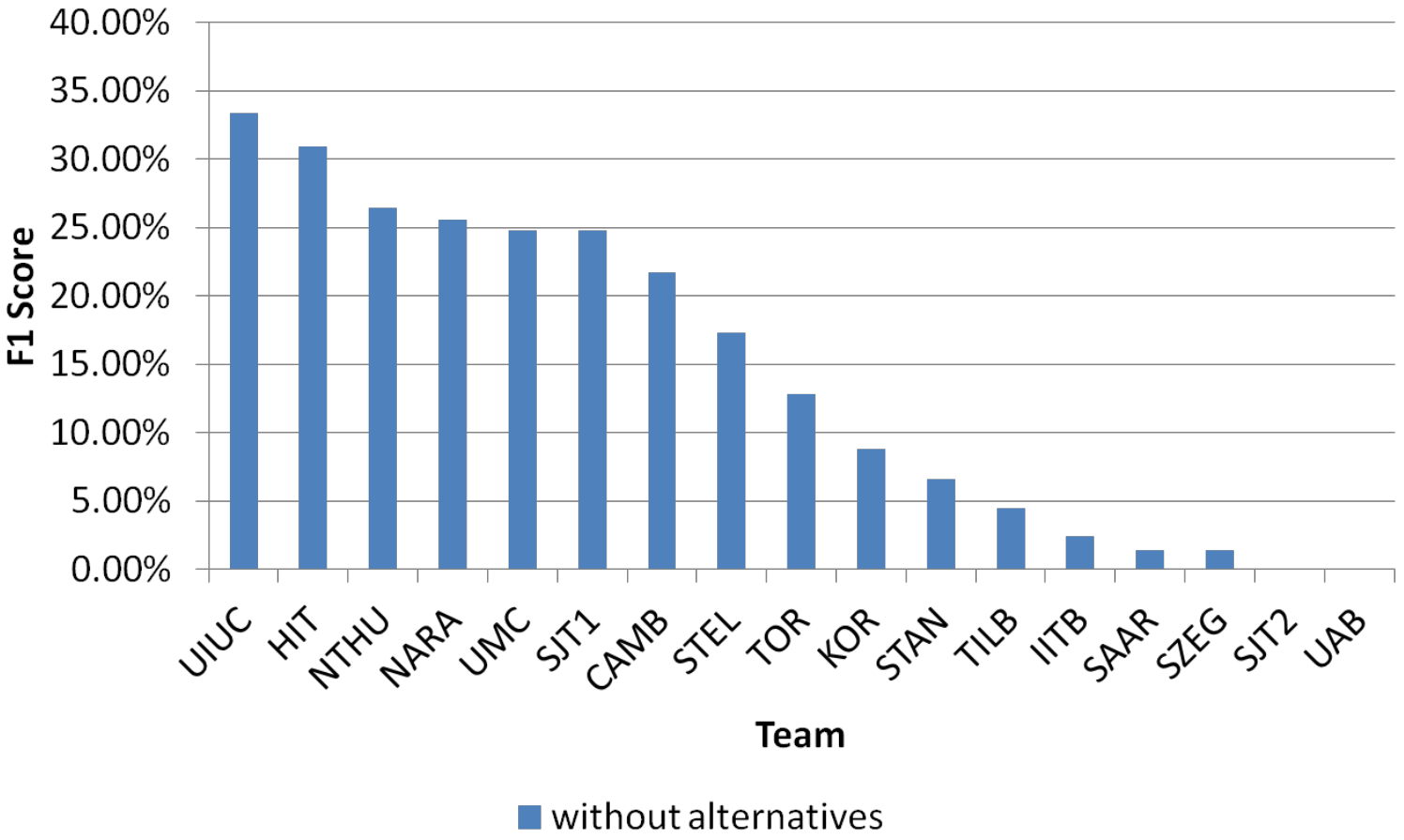
# System Scores without Alternative Answers



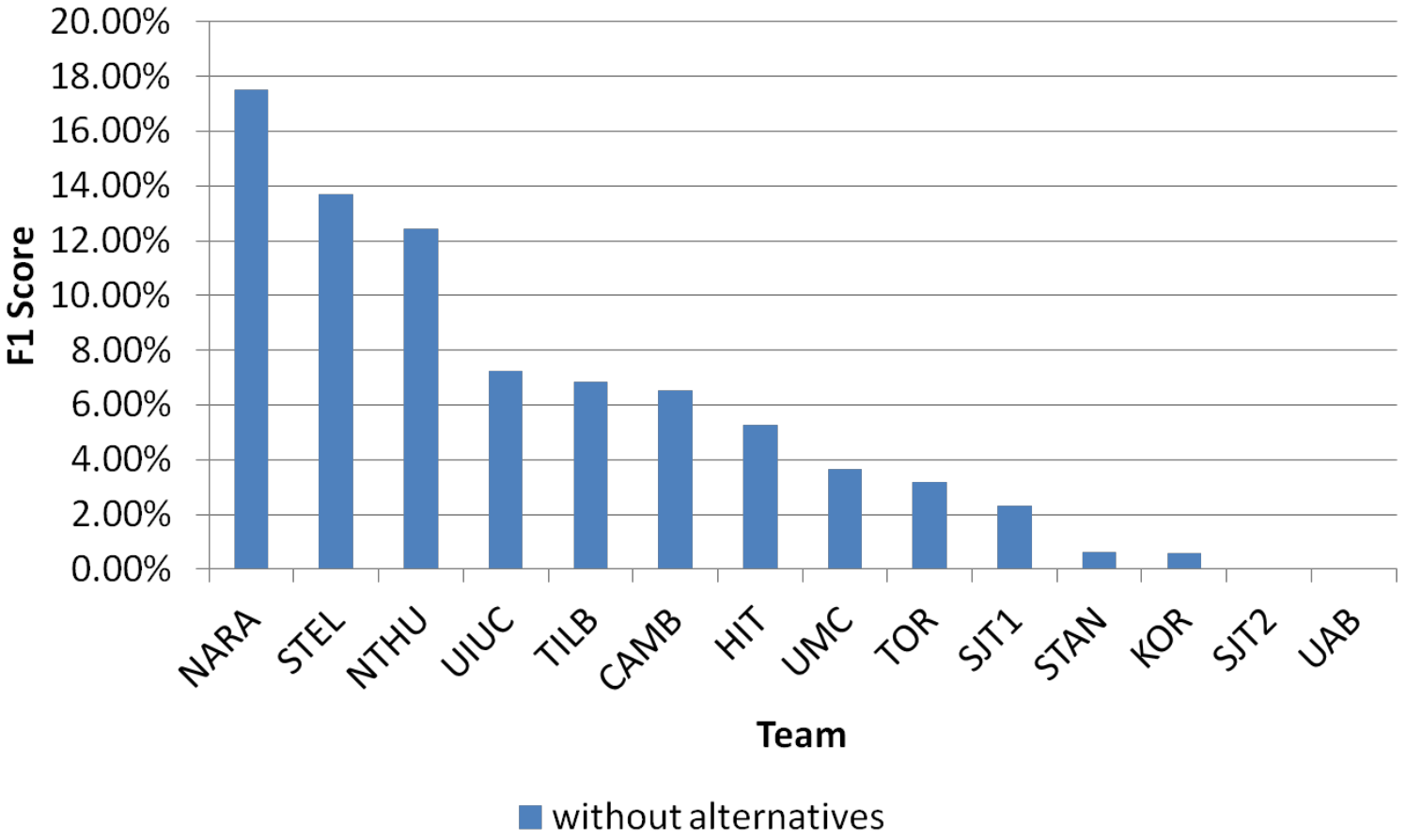
# System Scores with Alternative Answers



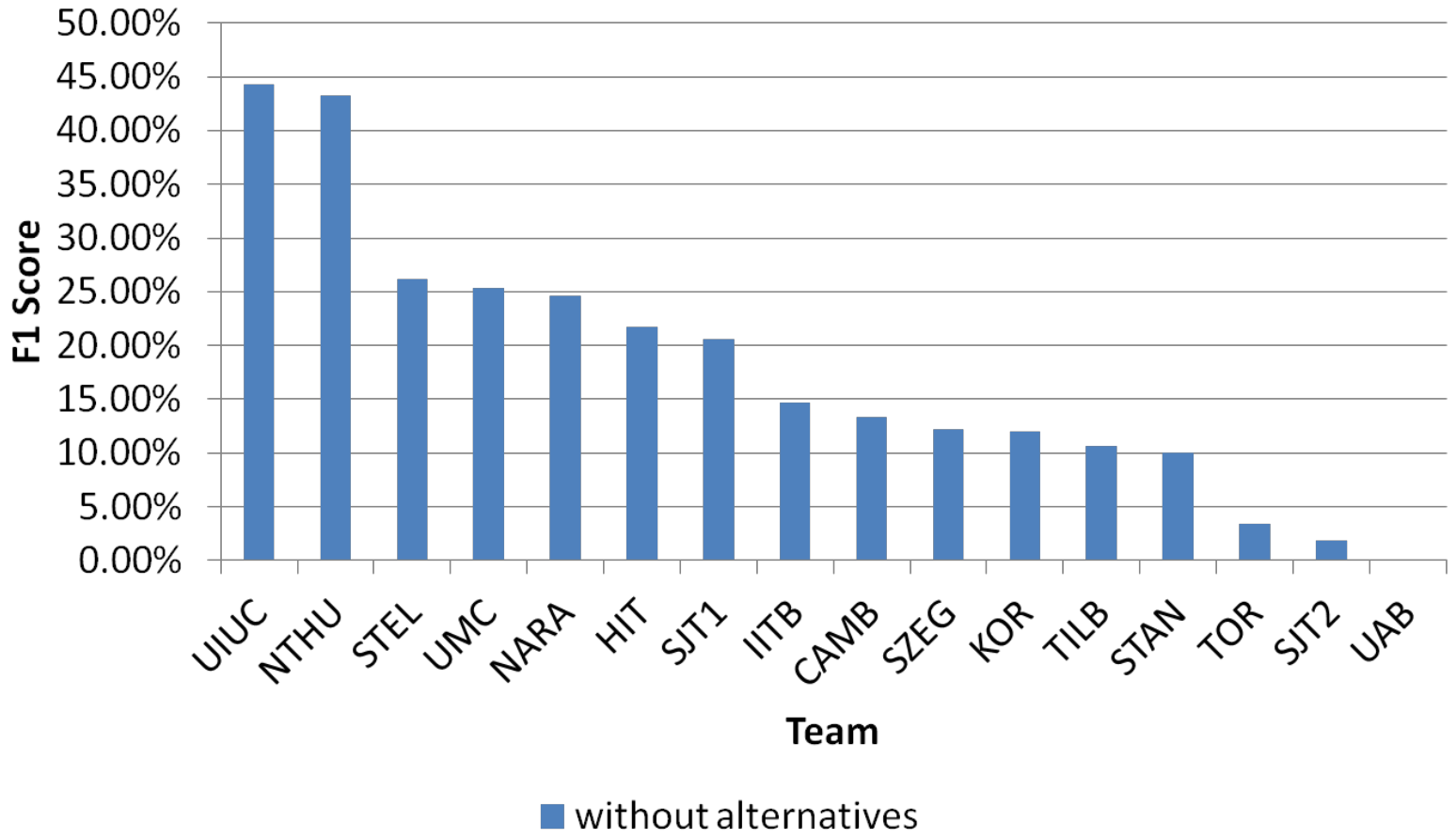
# Article/Determiner Errors



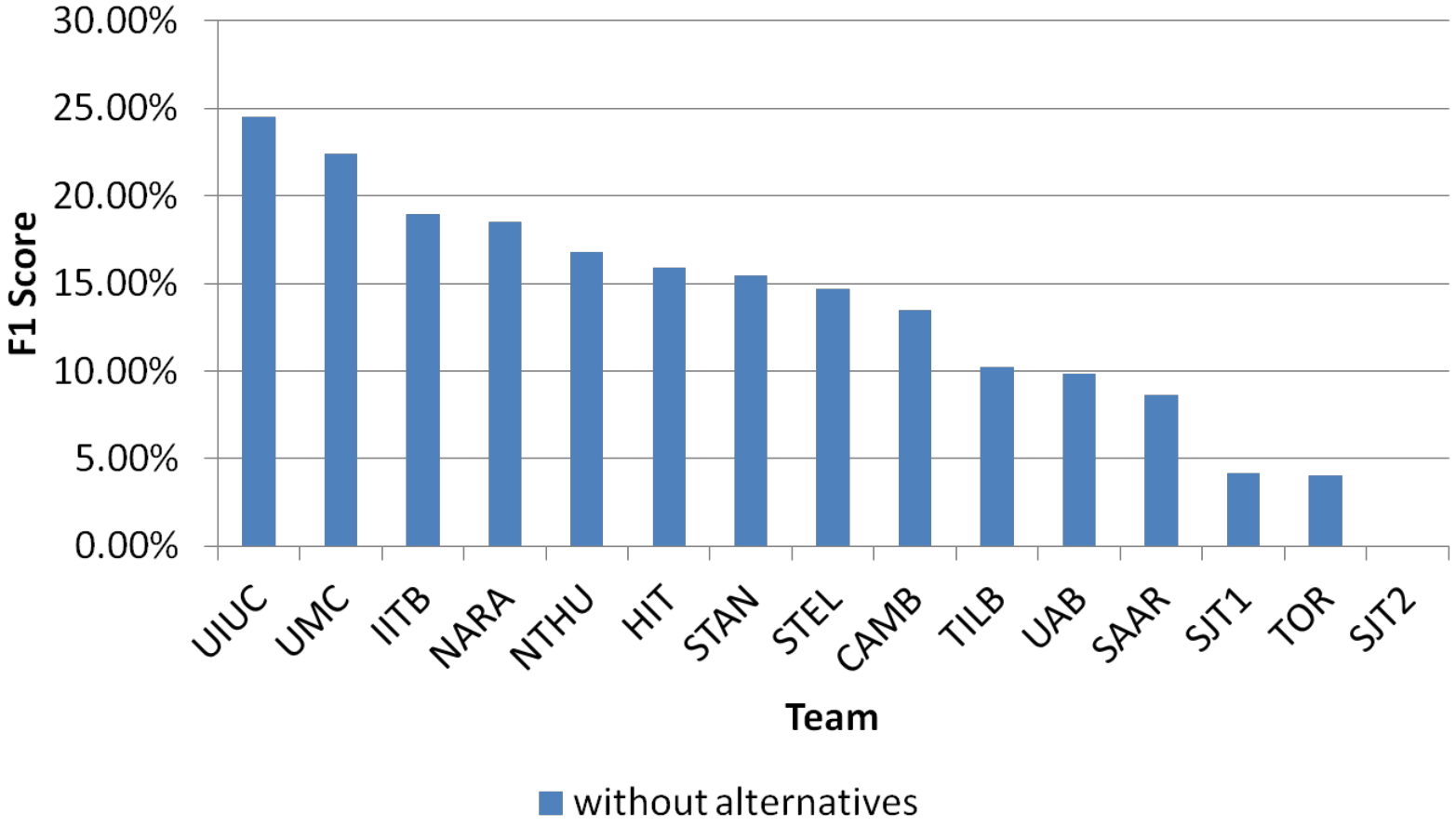
# Preposition Errors



# Noun Number Errors



# Verb Form/Subject–Verb Agreement Errors



# Extensions added in CoNLL-2013 Shared Task

- Expanded set of error types
  - Noun number, verb form, subject-verb agreement
- Fix the scoring anomaly with HOO scorer
- Test data freely available for future comparative evaluation

# Approaches

- A great variety of approaches
- Modeled as a classification task
  - One classifier per error type, e.g.,
    - Article: noun phrase → a/an, the, ϵ
    - Noun number: noun → singular, plural
  - Classifier can be:
    - Handcrafted rules
    - Learned from examples
    - Hybrid
- 11 teams adopted this approach

# Approaches

- Modeled as machine translation
  - Translate from “bad English” to “good English”
  - Cambridge, Stellenbosch, Toronto
- Language modeling approach
  - National Tsing Hua University
- Combination of learned classifier, machine translation, and language modeling
  - Nara Institute of Science & Technology

# Linguistic Features

- Lexical features (words, collocations, n-grams)
- Parts-of-speech
- Constituency parses
- Dependency parses
- Semantic features (semantic role labels)

# External Resources

- Cambridge Learner Corpus
- CMU Pronouncing Dictionary
- Europarl
- Gigaword
- Google Web 1T
- Lang-8
- Longman Dictionary
- Penn Treebank
- Wikipedia
- Wiktionary
- WordNet
- …

# A Beam-Search Decoder for GEC

- Dahlmeier & Ng, EMNLP 2012

- Grammatical error correction: viewed as translation (decoding) from "bad English" to "good English"
- Hypothesis (h): a revised sentence with one additional correction (edit) made

- Beam-search decoder:

While beam not empty & not max iterations do
  Propose new hypotheses // proposers
  Compute expert scores // experts
  Compute overall hypotheses scores // decoder model
  Prune hypotheses in beam

# A Beam-Search Decoder for GEC

- Proposers
  - Generate new hypotheses by making an incremental change (one correction/edit)
- Experts
  - Score hypotheses on particular aspects of grammaticality
- Decoder model
  - Combine evidence from experts into an overall score for each hypothesis
- A modular architecture that allows easy addition of new error types

# A Beam-Search Decoder for GEC

- Proposers
  - Article: Change the article (a/an, the, empty article ϵ) of each noun phrase (NP)
  - Preposition: Change the preposition of each prepositional phrase (PP)
  - Noun number: Change singular to plural noun or vice versa

# A Beam-Search Decoder for GEC

- Experts
  - Language model expert: compute the normalized n-gram language model score of a hypothesis
  - Article expert: compute the score of the article chosen for an NP in a hypothesis
  - Preposition expert: compute the score of the preposition chosen for a PP in a hypothesis
  - Noun number expert: compute the score of the noun form (singular/plural) chosen for a noun in a hypothesis
- Article/Preposition/Noun number expert is a linear classifier based on features like n-grams, POS tags, chunks, web-scale N-gram counts, & dependency parse trees in the neighboring context of an article/preposition/noun

# A Beam-Search Decoder for GEC

Decoder model

- Compute features of each hypothesis h:
  - Language model expert:
    - $score_{lm} = \frac{1}{|h|} logP(h)$
  - Article/preposition/noun number expert:
    - Average score: $score_{avg} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{u}^T f\left({\boldsymbol{x}_i}^h, {y_i}^h\right)\right)$
    - Delta score:$score_{delta} = \max_{i,y}\left(\boldsymbol{u}^T f\left({\boldsymbol{x}_i}^h, y\right) - \boldsymbol{u}^T f\left({\boldsymbol{x}_i}^h, {y_i}^h\right)\right)$
  - Correction count features ($count$)
    - Count how often each correction has been made in h

# A Beam-Search Decoder for GEC

- Decoder model
  - $g(h) = \begin{pmatrix} score_{lm} \\ score_{avg} \\ score_{delta} \\ count \\ \vdots \end{pmatrix}$
  - Linear combination of features of h into an overall score $s = w^T g\ (h)$
  - Optimize weight vector w with PRO (pairwise ranking optimization) to maximize F1 score on development set

# A Beam-Search Decoder for GEC

Evaluation on HOO 2011 test set (Dahlmeier & Ng, EMNLP 2012):

| System | F1 (M2 scorer) |
|---|---|
| UIUC (top team) | 17.59 |
| Pipeline | 20.67 |
| Beam-search decoder | 23.48 |

Evaluation on CoNLL-2013 test set also shows that the beam-search decoder performs as well as the best participating system

# Open Research Issues

- Much work remains to be done!
  - State-of-the-art performance: 31% recall, 62% precision
- Statistical approaches have potential to significantly outperform a hand-crafted rule-based approach
  - "Big Data" movement: Exploit very large corpora
    - To learn a language well, we need to be exposed to the language
  - a la statistical machine translation (SMT outperforms hand-crafted rule-based MT)

# Open Research Issues

- Expand the error types dealt with
- Efficiently search for the best corrections
- Upper bound of human agreement
  - Far from 100%
  - Not all errors are equal
- Trade-off between precision and recall
- …

# Conclusion

- Resurgence of a somewhat neglected field
- Performance of grammatical error correction may see significant improvements in the near future
- A difficult task that has far-reaching real-world impact