Automated Grammatical Error Correction: The State of the Art

Hwee Tou Ng

Department of Computer Science
School of Computing
National University of Singapore

8 Dec 2014

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, linguistic engineering approach
  - Limited coverage (detects none of the 5 sample grammatical errors shown)

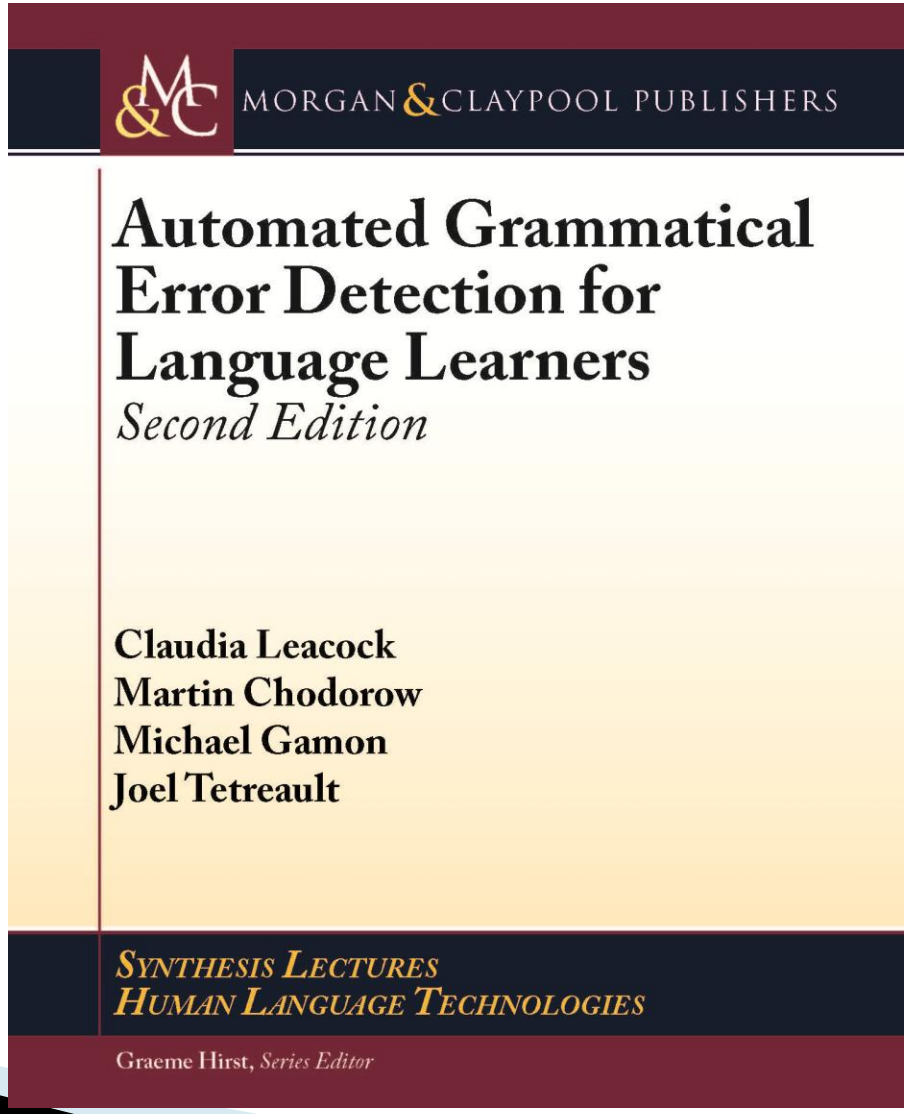
Current Landscape

- Commercial software available:

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, 2014

Introductory Book (2014)



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
- Four 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)
  - CoNLL 2014 Shared Task (Ng et al., 2014)

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
- Handbook of Automated Essay Evaluation: Current Applications and New Directions, Shermis and Burstein (ed), 2013
- 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

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
- Evaluation metric: F1
- One human annotator provided the gold-standard annotations

CoNLL-2014 shared task

- 5 error types → all (28) error types
- Evaluation metric F_1 → $F_{0.5}$ (emphasize precision over recall)
- One → Two human annotators (who independently annotated the test essays)

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
- 28 error types

NUCLE Error Types (Version 3.2)

Error Tag	Error Type	Error Tag	Error Type
Vt	Verb tense	Srun	Runons, comma splices
Vm	Verb modal	Smod	Dangling modifiers
V0	Missing verb	Spar	Parallelism
Vform	Verb form	Sfrag	Sentence fragment
SVA	Subject–verb agreement	Ssub	Subordinate clause
ArtOrDet	Article or determiner	WOinc	Incorrect word order
Nn	Noun number	WOadv	Incorrect adj/adv order
Npos	Noun possessive	Trans	Link words/phrases
Pform	Pronoun form	Mec	Punctuation, capitalization, spelling, etc
Pref	Pronoun reference	Rloc–	Redundancy
Wci	Wrong collocation/idiom	Prep	Preposition
Wa	Acronyms	Cit	Citation
Wform	Word form	Others	Other errors
Wtone	Tone (formal/informal)	Um	Unclear meaning

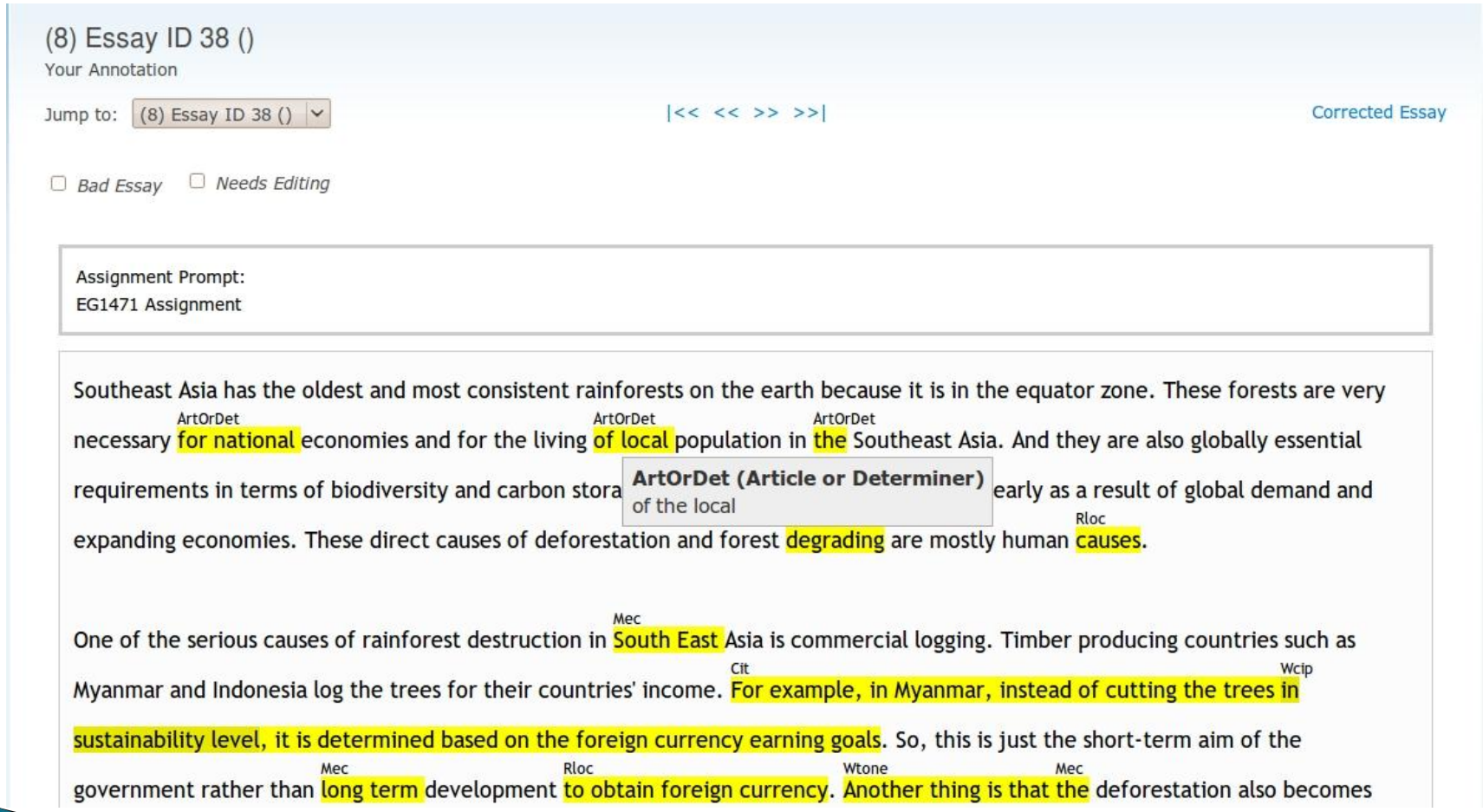
Usage of Training Data and Tools

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

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
Rloc
expanding economies. These direct causes of deforestation and forest degrading are mostly human 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

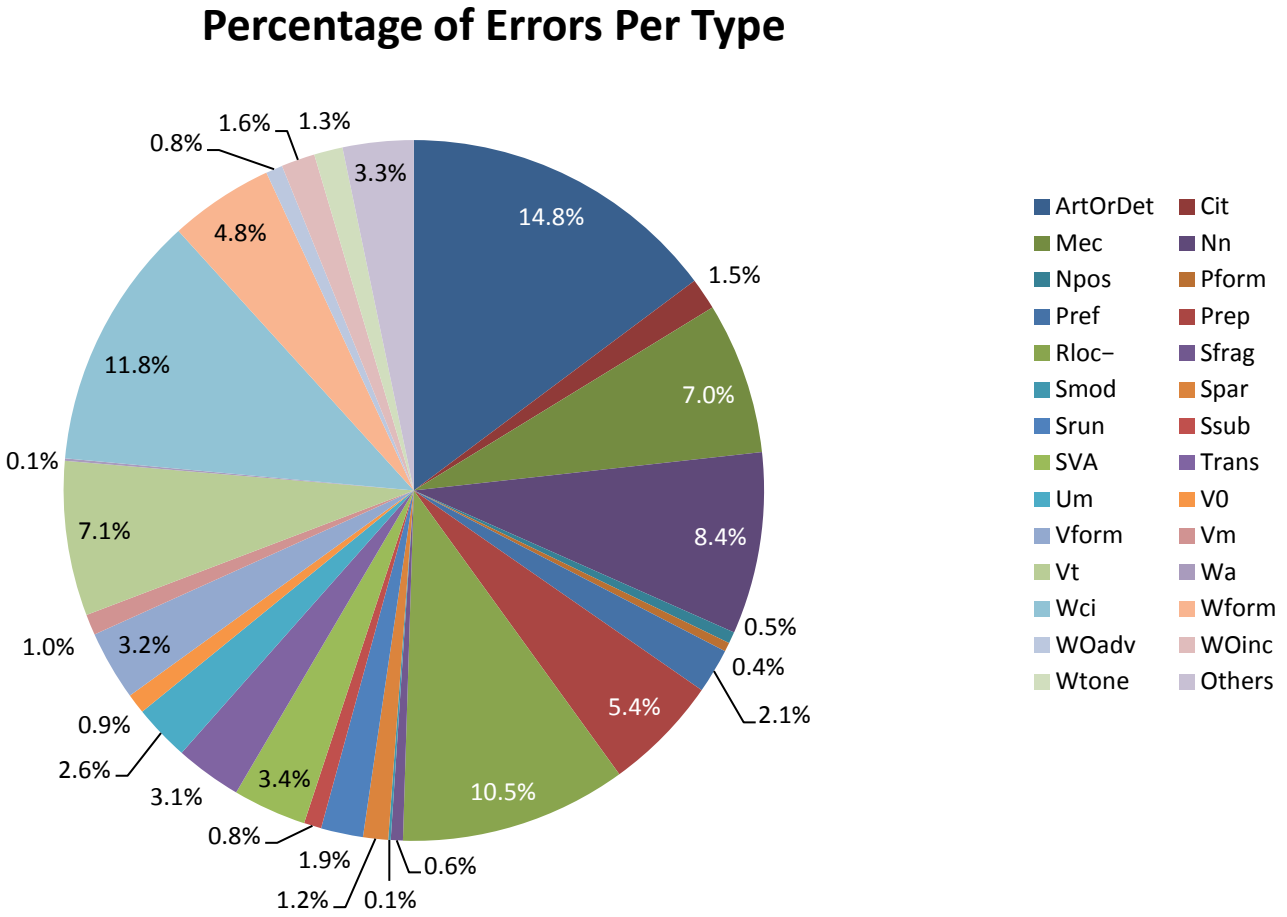
```
<MISTAKE start_par="0" start_off="5" end_par="0" end_off="9">
<TYPE>ArtOrDet</TYPE>
<CORRECTION>the past</CORRECTION>
</MISTAKE>
```

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

Statistics of NUCLE (version 3.2)

- # essays = 1,397
- # sentences = 57,151
- # word tokens = 1,161,567
- # errors (in all 28 error types) = 44,912

Statistics of Errors in NUCLE



Test Data for CoNLL-2014

- 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 = 1,312
- # word tokens = 30,144

Two Prompts for the Test Essays

- “The decision to undergo genetic testing can only be made by the individual at risk for a disorder. Once a test has been conducted and the results are known, however, a new, family-related ethical dilemma is born: Should a carrier of a known genetic risk be obligated to tell his or her relatives?” Respond to the question above, supporting your argument with concrete examples.
- While social media sites such as Twitter and Facebook can connect us closely to people in many parts of the world, some argue that the reduction in face-to-face human contact affects interpersonal skills. Explain the advantages and disadvantages of using social media in your daily life/society.

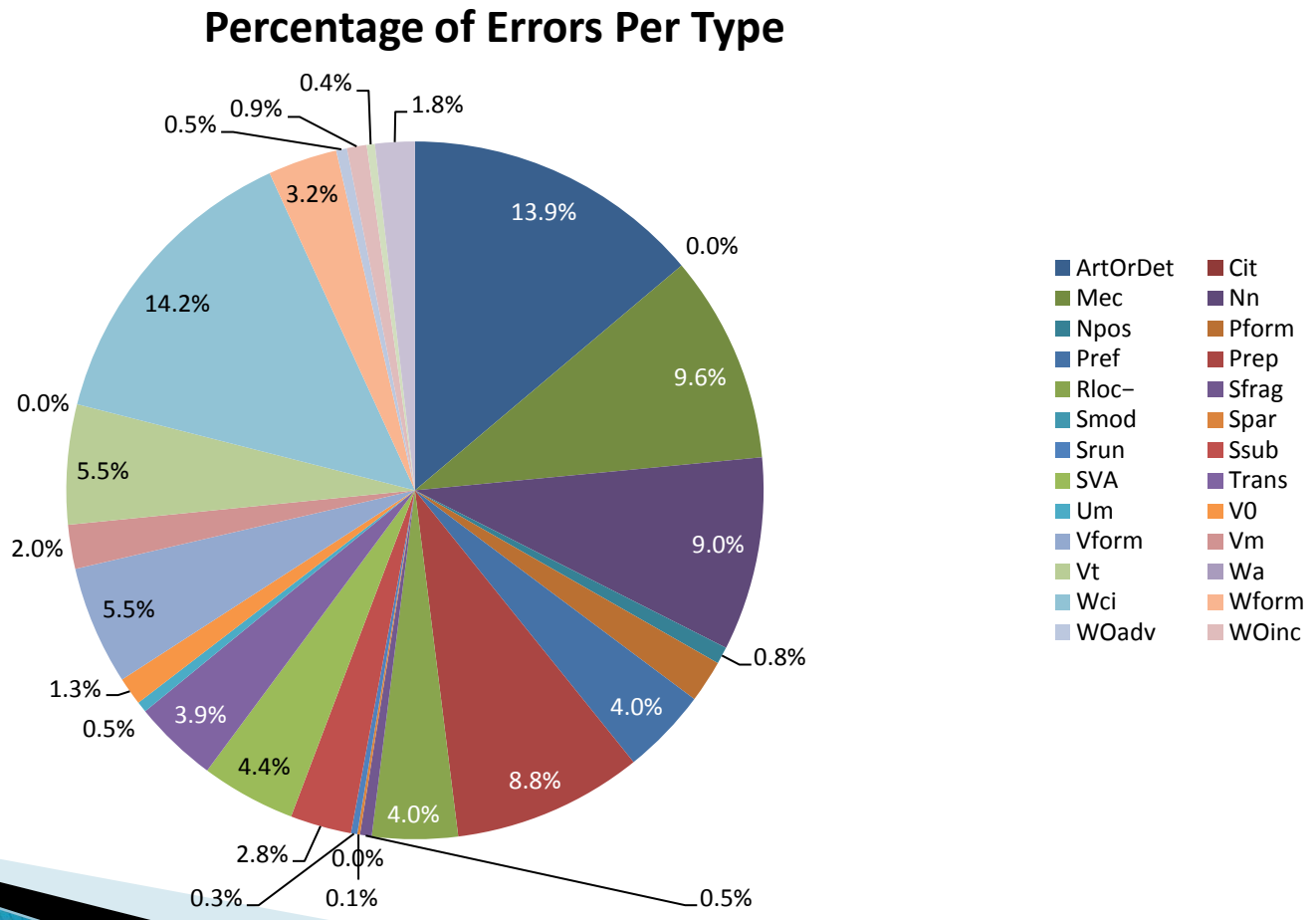
Test Data for CoNLL-2014

- Annotation on test essays carried out independently by two native speakers of English
- Test essays and annotations freely available at the shared task home page:

http://www.comp.nus.edu.sg/~nlp/conll14st.html

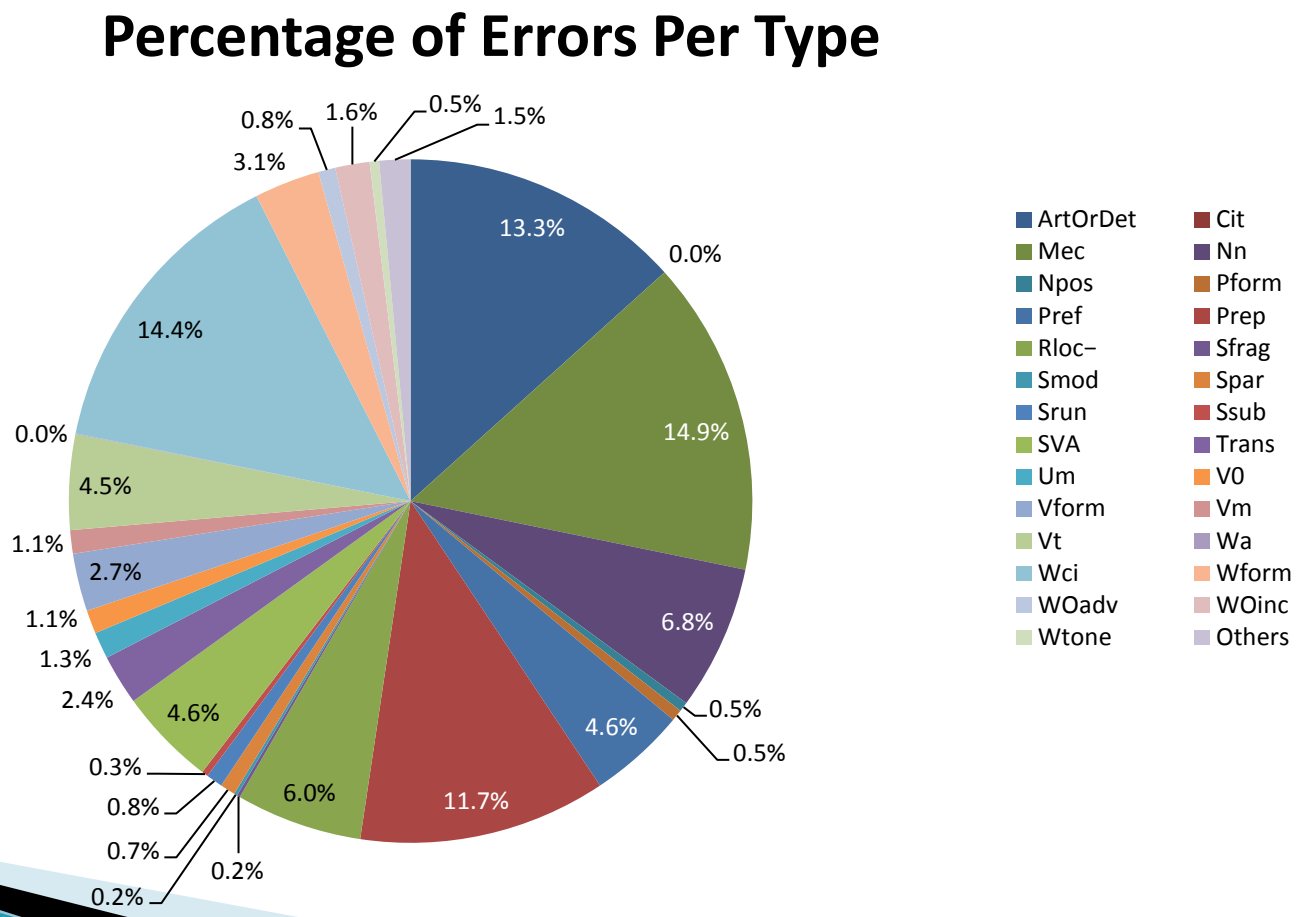
Statistics of Errors in CoNLL-2014 Test Data (Annotator 1)

Five error types of CoNLL-2013 account for 41.6% of all errors



Statistics of Errors in CoNLL-2014 Test Data (Annotator 2)

Five error types of CoNLL-2013 account for 39.1% of all errors



Evaluation

- Edits: corrections
- How well the proposed system edits (e_i) match the gold-standard edits (g_i)
- Recall (R), Precision (P), $F_{0.5}$ measure (emphasize precision over recall)
- n = # of sentences

$$R = \frac{\sum_{i=1}^{n} |\mathbf{g}_i \cap \mathbf{e}_i|}{\sum_{i=1}^{n} |\mathbf{g}_i|} \qquad P = \frac{\sum_{i=1}^{n} |\mathbf{g}_i \cap \mathbf{e}_i|}{\sum_{i=1}^{n} |\mathbf{e}_i|} \qquad F_{0.5} = \frac{(1+0.5^2) \times R \times P}{R + 0.5^2 \times P}$$

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 by a system:
    - There is no doubt , tracking system has brought many benefits .
  - System edits e = { a doubt → doubt }
  - R = 1/3, P = 1/1
  - $F_{0.5} = 1.25 \times 1/3 \times 1 \ / \ (1/3 + 0.25 \times 1) = 5/7$

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 by a system:
  - There is no doubt , tracking system has brought many benefits .
- GNU wdiff gives system edits e = { a → ϵ }
- HOO scorer gives erroneous scores: $R = P = F_{0.5} = 0$

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
- Overcome scoring anomaly of HOO scorer
- Available from the shared task home page:

http://www.comp.nus.edu.sg/~nlp/conll14st.html

CoNLL-2013 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

CoNLL-2014 Participating Teams (13)

Team ID	Affiliation
AMU	Adam Mickiewicz University
CAMB	University of Cambridge
CUUI	Columbia University and the University of Illinois at Urbana-Champaign
IITB*	Indian Institute of Technology, Bombay
IPN*	Instituto Politécnico Nacional
NARA	Nara Institute of Science and Technology
NTHU	National Tsing Hua University
PKU*	Peking University
POST	Pohang University of Science and Technology
RAC	Research Institute for Artificial Intelligence, Romanian Academy
SJTU	Shanghai Jiao Tong University
UFC*	University of Franche-Comté
UMC	University of Macau

*:Teams that submitted after the submission deadline

Asia: 7
Europe: 4
North America: 2

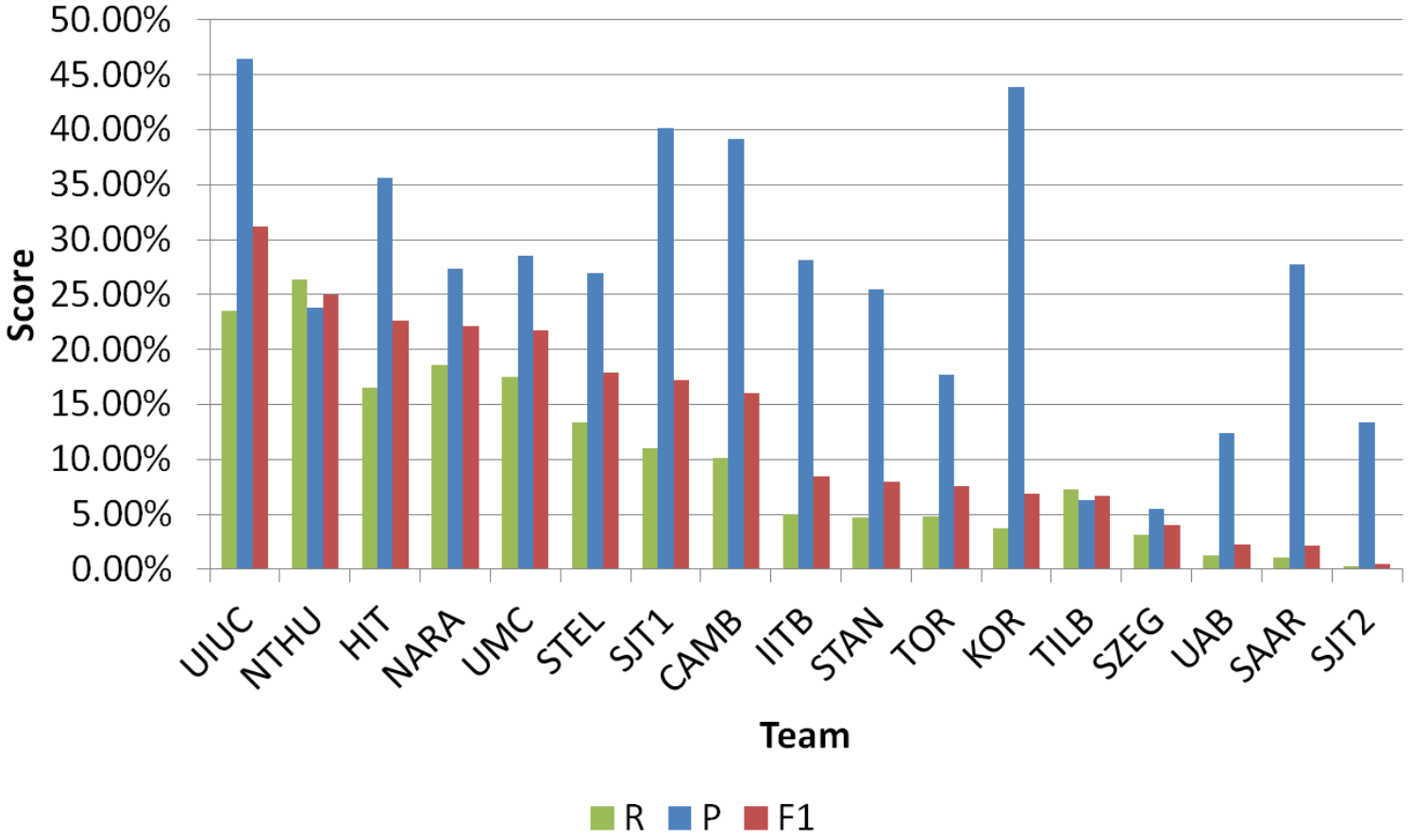
Alternative Annotations

- 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
- Used in HOO 2011, HOO 2012, CoNLL 2013, & CoNLL 2014
- Extend M2 scorer to deal with multiple alternative gold-standard annotations

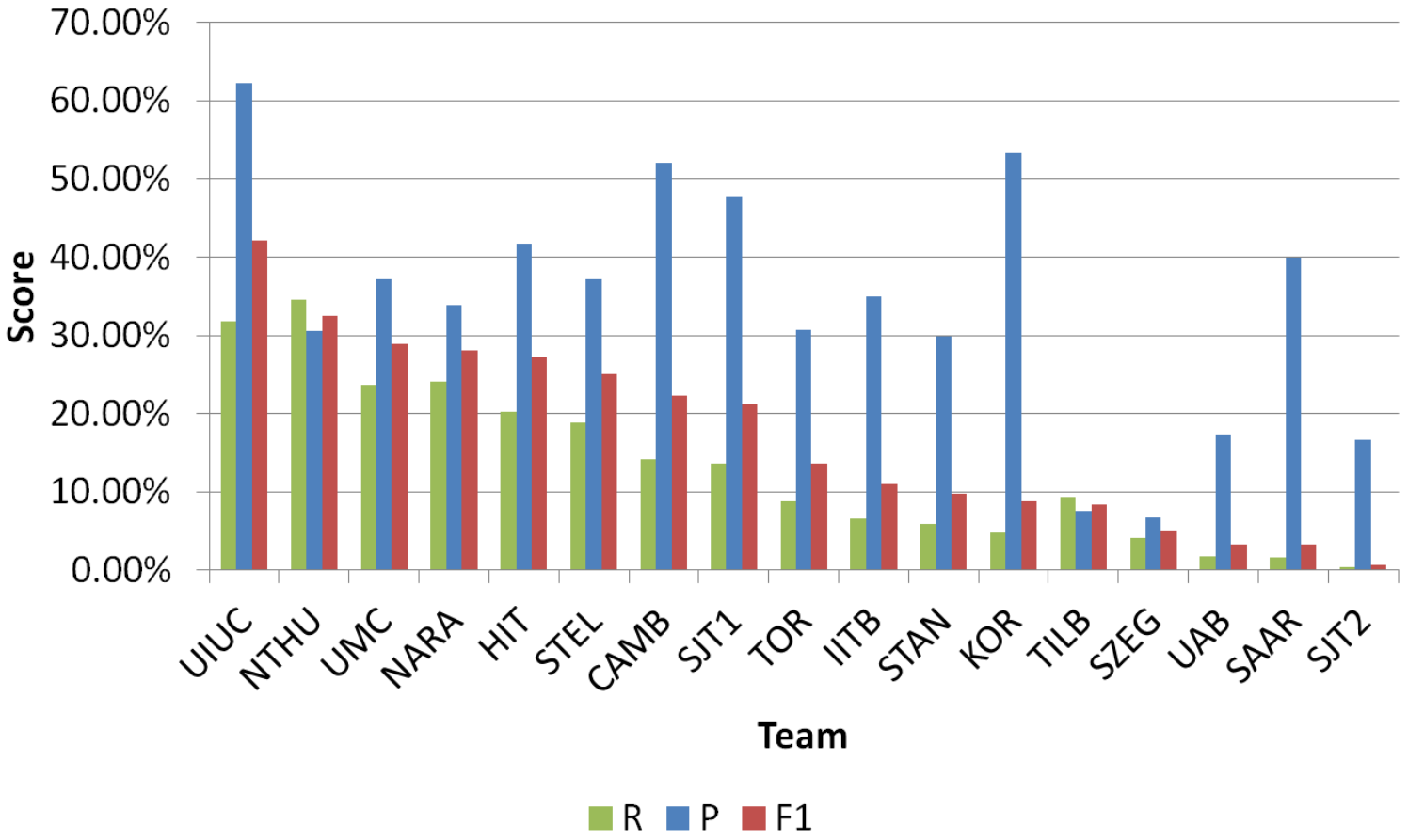
Alternative Annotations

- Five teams (NTHU, STEL, TOR, UIUC, UMC) submitted alternative answers in CoNLL 2013
- Three teams (CAMB, CUUI, UMC) submitted alternative answers in CoNLL 2014
- Alternative answers proposed were judged by the same annotators who provided the original gold-standard annotations
- F1 / $F_{0.5}$ scores of all teams improve when evaluated with alternative answers
- For future research which uses the test data, we recommend reporting scores in the setting that does *not* use alternative answers

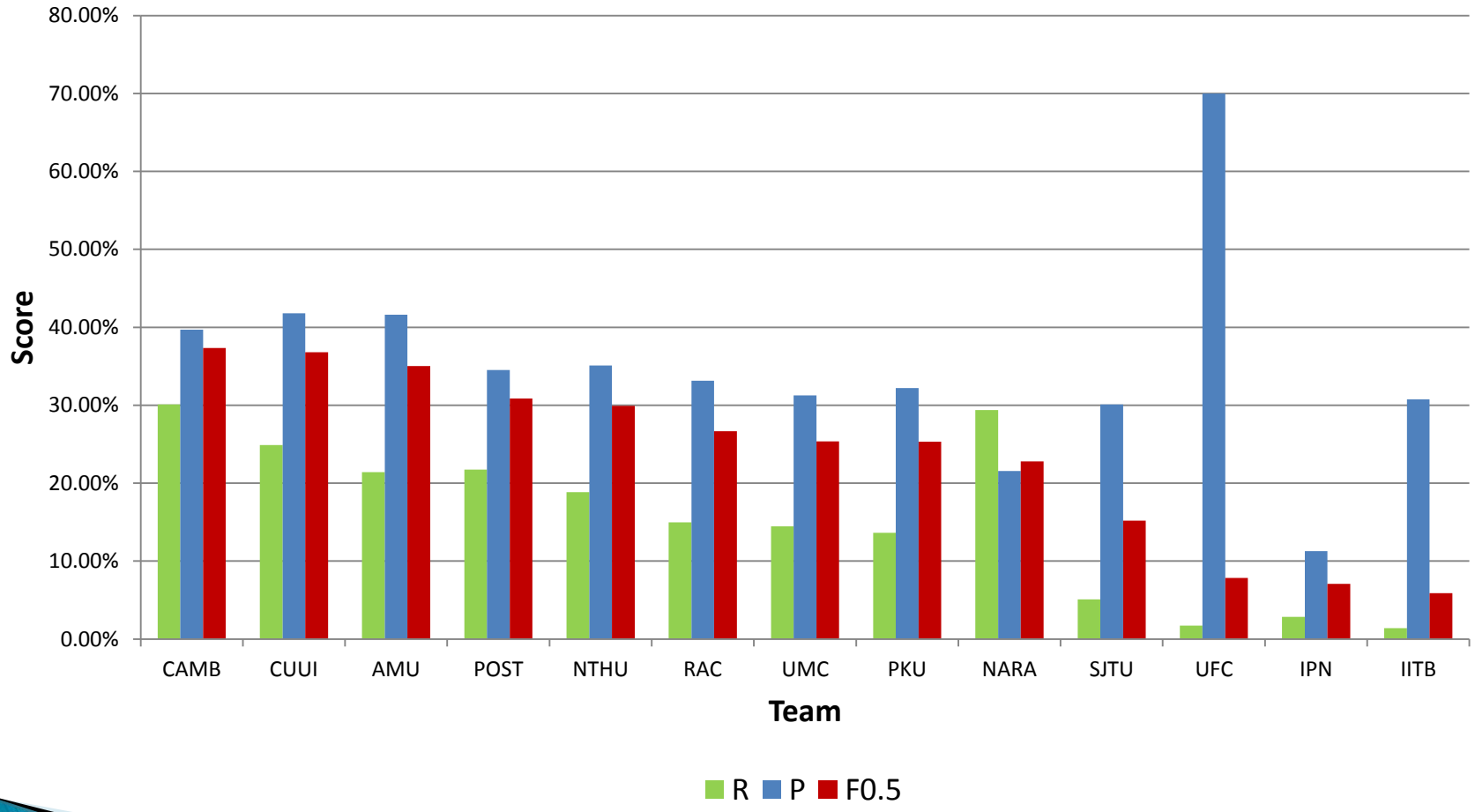
CoNLL-2013 System Scores without Alternative Answers



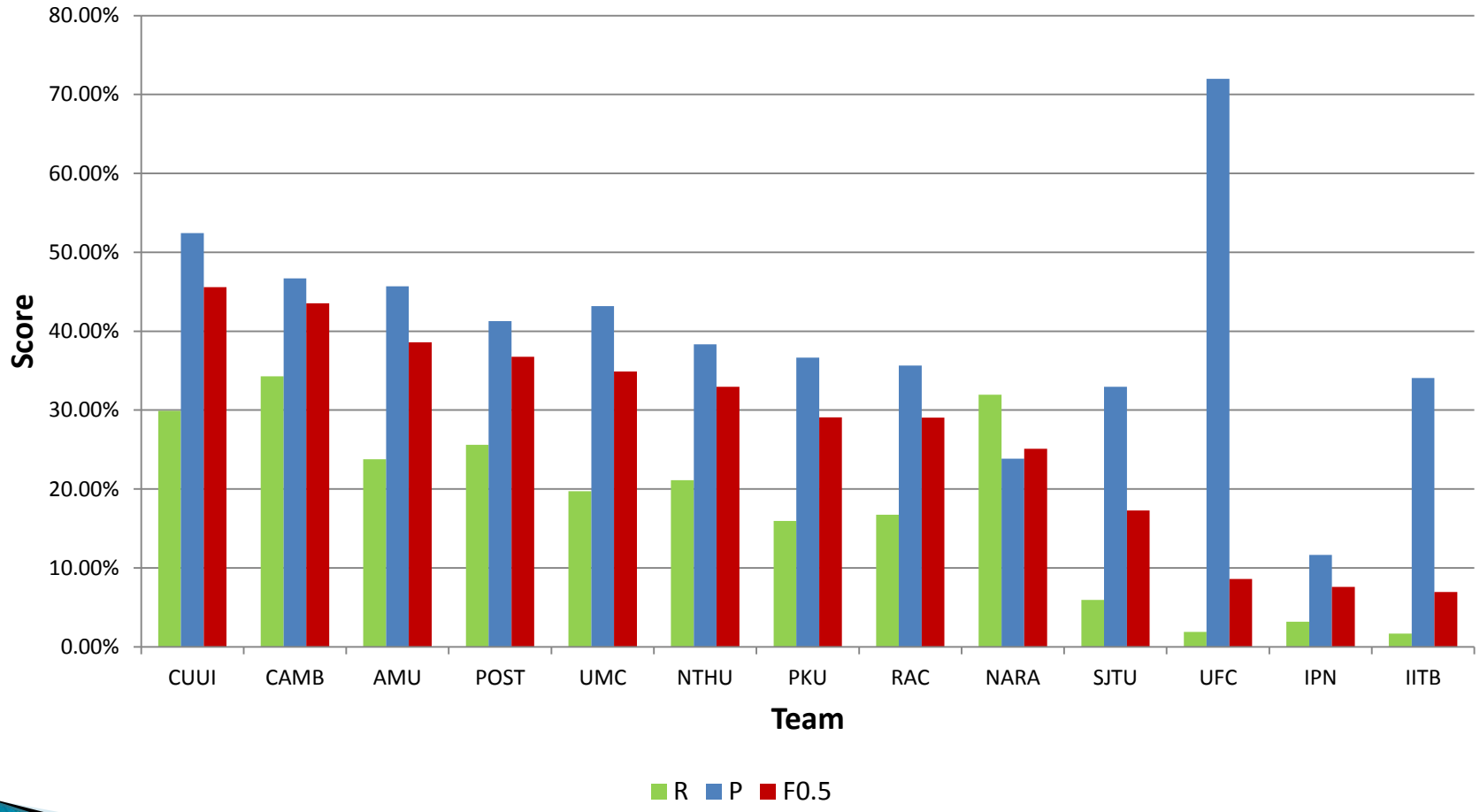
CoNLL-2013 System Scores with Alternative Answers



CoNLL-2014 System Scores without Alternative Answers



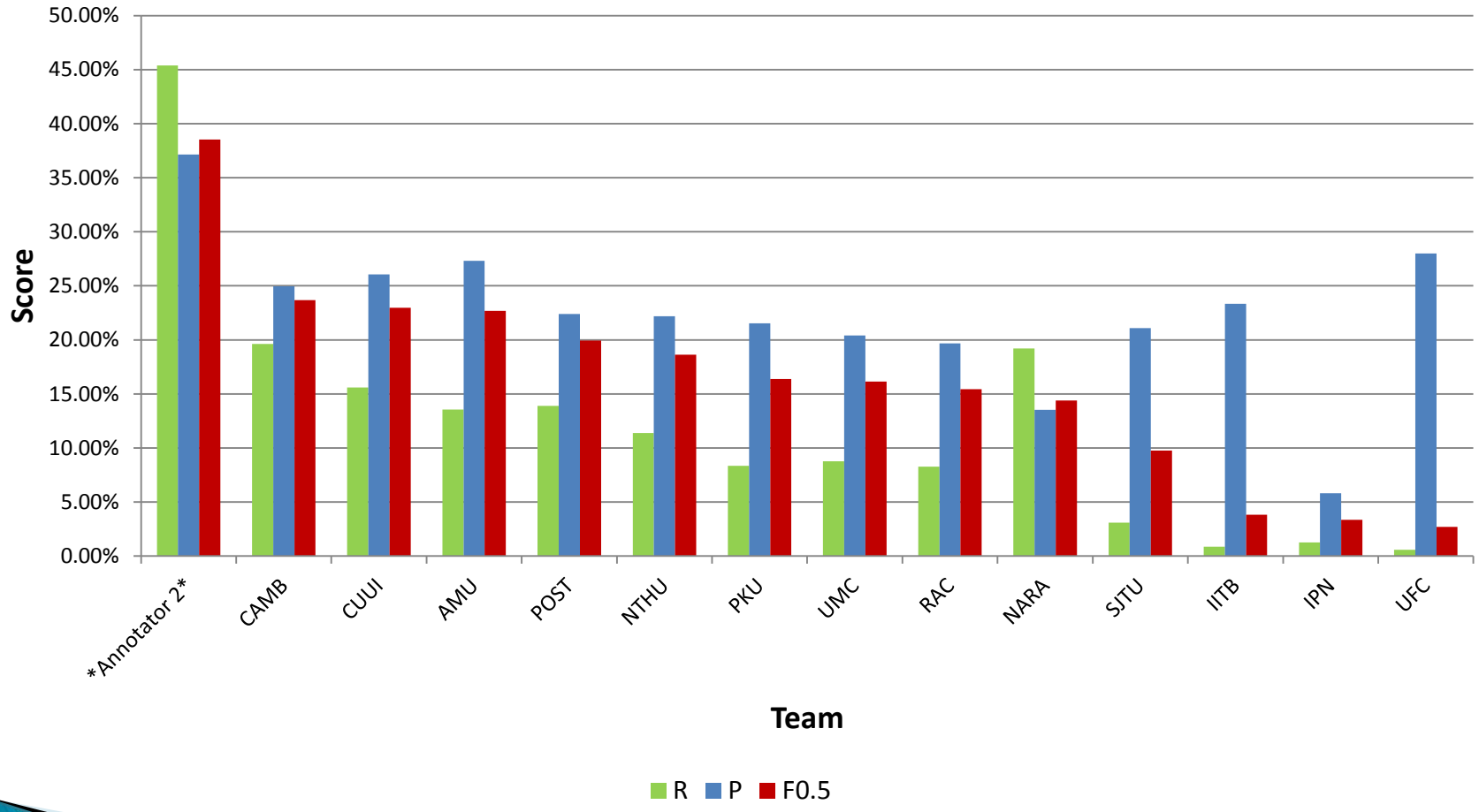
CoNLL-2014 System Scores with Alternative Answers



Cross Annotator Comparison

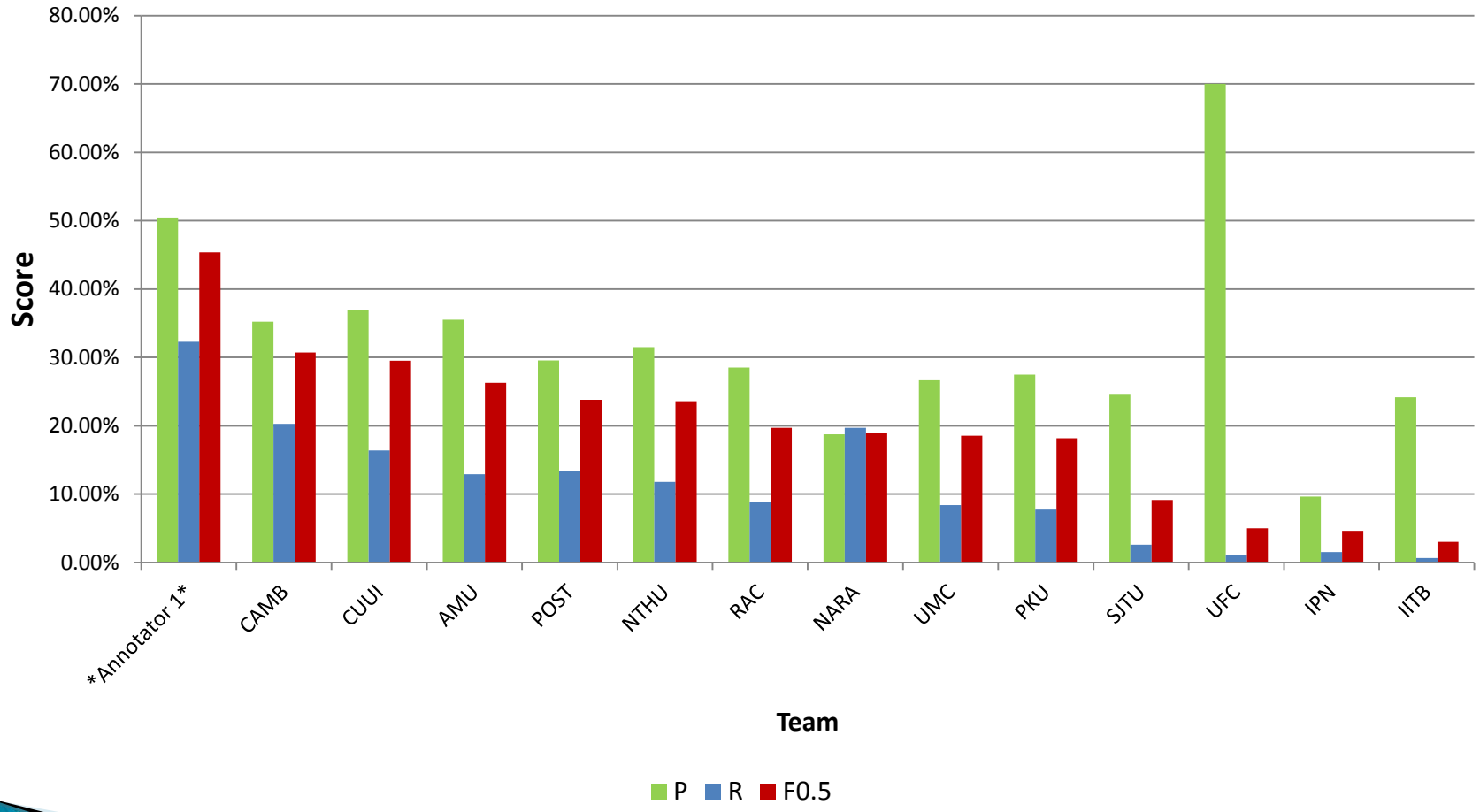
- Kappa coefficient for identification = 0.43
- Measures the extent to which the two annotators agreed which words needed correction and which did not (regardless of the error type or correction)
- Moderate agreement (0.40 – 0.60)

System Scores Based on Annotator 1



61% of human $F_{0.5}$ score

System Scores Based on Annotator 2



67% of human $F_{0.5}$ score

Linguistic Knowledge

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

External Resources

- Academic Word List
- Aspell
- British National Corpus
- Cambridge Learner Corpus
- Cambridge "Write and Improve" SAT system
- CommonCrawl
- CoNLL-2013 test set
- English Vocabulary Profile corpus
- Europarl
- First Certificate in English (FCE) corpus
- Gigaword
- GingerIt
- Google Books Syntactic N-grams
- Google Web 1T
- Lang-8
- Lucene Spellchecker
- Microsoft Web LM
- Wikipedia
- …

Approaches to Grammatical Error Correction

- Two dominant approaches:
  - **Classification approach**
  - **Translation approach**

Classification Approach

- 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
  - CUUI system (Columbia U – UIUC)

The CUUI System

- Each error type is handled by an independent classifier
- A confusion set of classes per classifier (multi-class classification task)
- A confusable word instance w → A vector of features derived from a context window around w
- Features: Rely on a POS tagger and a chunker

The CUUI System

- Does not deal with word choice error type (WCI)
- Error types dealt with: ArtOrDet, Prep, Nn, SVA, Vform, Wform, Mec, Wtone
- Training data: learner data (NUCLE) and/or native data (Google Web 1T 5-gram)
- Learning algorithms: averaged perceptron, naïve Bayes
- Pattern/rule-based method for Wtone errors
- *Pipeline* system of applying classifiers, one after another (ArtOrDet, Prep, Nn, SVA, Vform, Wform, Mec, Wtone)

The CUUI System

- Model combination
  - Combine two models:
    - Averaged perceptron trained on learner data (NUCLE) with richer features (POS tags, dependency parse features, source word of the author)
    - Naïve Bayes trained on native data (Google Web 1T 5-gram)

The CUUI System

- Joint inference
  - Prevent inconsistent predictions for interacting errors (e.g., noun number and subject-verb agreement)
  - Global inference via Integer Linear Programming

Classification Approach

- Advantages:
  - Able to focus on each individual error type using a separate classifier
- Disadvantages:
  - Complicate the design since we need to build many classifiers
  - Need additional mechanism to deal with multiple interacting error types

Translation Approach

- Modeled as statistical machine translation (SMT)
  - Translate from “bad English” to “good English”
  - Do not target specific error types, but rather generic text transformation
  - Cambridge, AMU systems
  - Give state-of-the-art performance in CoNLL 2014 shared task

The AMU System

- Adam Mickiewicz University, Poland
- Phrase-based statistical machine translation (SMT)
- Make use of large scale error-corrected texts
- Lang-8: Social language learners' platform http://lang-8.com/
- Early SMT approach: correct countability errors for mass nouns (Brockett, Dolan, Gamon, 2006)

The AMU System

- No reordering models
- Translation model trained from "parallel" texts:
  - NUCLE
  - Lang-8 corpus: 3.7 million sentence pairs, 51.2 million tokens (uncorrected source side)
- Large language models (LM)
  - 3-gram LM estimated from English Wikipedia (3.2×10^9 tokens)
  - 5-gram LM estimated from CommonCrawl data (4.4×10^{11} tokens)
- Part of NUCLE is used as the tuning data
- Tuning based on the $F_{0.5}$ metric computed by the M2 scorer

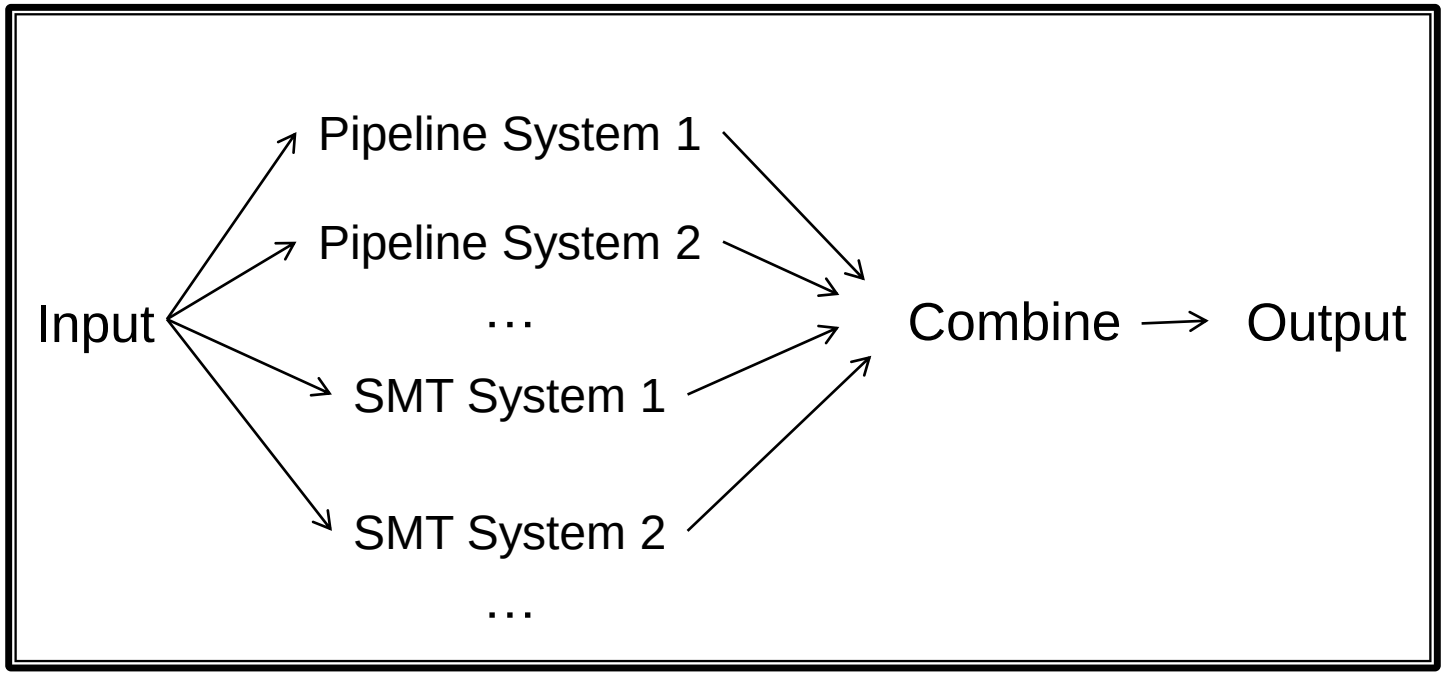
Translation Approach

- Advantages:
  - Naturally take care of interaction among multiple error types
  - Better coverage of different error types
- Disadvantages:
  - Rely on error-annotated learner texts, which are expensive to produce

System Combination Approach

- Idea: Combine the outputs of classification and SMT systems to produce an overall better output
- Best of both worlds:
  - Error type-specific classifiers + dealing with multiple interacting errors
- Susanto, Phandi, & Ng (EMNLP 2014)

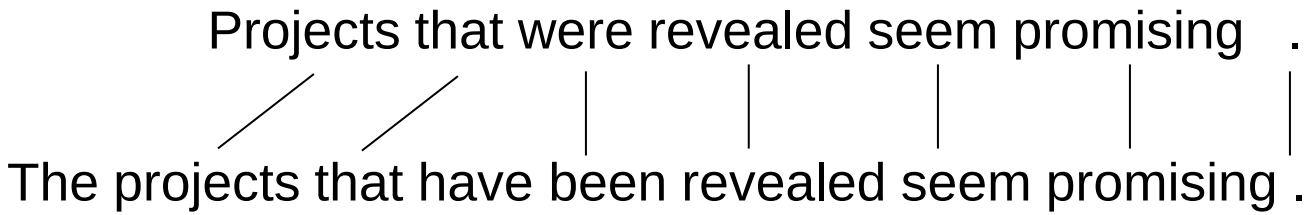
System Combination Approach



Combine: MEMT (Multi-Engine Machine Translation) system combination approach of Heafield & Lavie (2010)

MEMT Combination Scheme

- Step 1: Alignment
  - Run METEOR aligner on every pair of system outputs for a given sentence
  - Allow case-insensitive exact matches, stem matches, synonyms, unigram paraphrases
  - Example:



MEMT Combination Scheme

- Step 2: Search
  - Beam search over the aligned sentences
  - Hypotheses are constructed as follows:
    - Append the first (leftmost) unused word from a system
    - Mark the appended word and those aligned with it as "used"
  - A hypothesis is scored based on a set of features (language model, n-gram match, length, backoff)

Component Systems

- Four individual error correction systems:
  - Two pipeline-of-classifiers systems
  - Two phrase-based SMT systems

Pipeline of Classifiers

- Confidence-weighted linear classifiers for correcting noun number, preposition, and article errors
- Rule-based classifiers for correcting punctuation, verb form, and SVA errors
- Dictionary-based spell checker

Pipeline of Classifiers

Step	Pipeline 1 (P1)	Pipeline 2 (P2)
1	Spelling	Spelling
2	Noun number	Article
3	Preposition	Preposition
4	Punctuation	Punctuation
5	Article	Noun number
6	Verb form, SVA	Verb form, SVA

Statistical Machine Translation (SMT)

- Phrase-based SMT systems built using Moses
- SMT 1 (S1): two phrase tables trained on NUCLE and Lang-8 separately
- SMT 2 (S2): a single phrase table trained on the concatenation of NUCLE and Lang-8

Data

- Training data
  - Error-annotated learner corpora:
    - NUCLE (1.16M source tokens)
    - Lang-8 (12.95M source tokens)
  - English Wikipedia (1.78B tokens)
- Development data
  - CoNLL-2013 test set (29K tokens)
- Test data
  - CoNLL-2014 test set (30K tokens)

Results

System	P	R	$F_{0.5}$
Pipeline			
P1	40.24	23.99	35.44
P2	39.93	22.77	34.70
SMT			
S1	57.90	14.16	35.80
S2	62.11	12.54	34.69
Combined			
P1+S1	53.85	17.65	38.19
P2+S2	56.92	16.22	37.90
P1+P2+S1+S2	53.55	19.14	39.39
Top 4 Systems in CoNLL-2014			
CAMB	39.71	30.10	37.33
CUUI	41.78	24.88	36.79
AMU	41.62	21.40	35.01
POST	34.51	21.73	30.88

Highest published $F_{0.5}$ score on CoNLL-2014 test set

Open Research Issues

- Much work remains to be done:
  - State-of-the-art performance: 61-67% of human performance
- Statistical approaches have potential to significantly outperform a hand-crafted, knowledge engineering approach
  - "Big Data" movement: Exploit very large corpora
    - To learn a language well, we need to be exposed to the language
  - Lang-8 data looks promising

Open Research Issues

- Upper bound of human agreement
  - Far from 100% based on current measurement
  - Not all errors are equal
- Trade-off between precision and recall
- Training data selection

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