Automatic Grammatical Error Correction for Language Learners

Joel Tetreault
Claudia Leacock
What is a grammatical error?

1. Syntax: “Each language has its own systematic ways through which words and sentences are assembled to convey meaning.” Fraser & Hodson (1978)
   - Syntax errors are rule-driven (e.g. subj-verb agreement) thus easier to learn

2. Usage: Conventional usage habits
   - A wrong preposition or missing determiner – do not break rules of syntax but of usage.
   - Usage errors are most common for learners – greater reliance on memory than rules
Focus on English

- Need
  - Over a billion people speak English as a second or foreign language worldwide
  - By 2025, estimated that English language learners will make up 25% of the US public school population
  - 725,000 international students at US universities
  - 13 million college students in China took the College English Test in 2006
  - 27 million people have taken the TOEFL

- Practical
  - English language has most resources
Goals of Tutorial

- Challenges that language learners face
- Challenges of designing tools to assist learners
- Brief History of GEC
- State-of-the-art approaches for different error types
First to use SL; explored 3 training paradigms: well-formed, artificial errors, real errors. Webscale Data Artificial Errors. Real Errors. 4 Shared Tasks 2011-2014. GEC takes off! 4 Shared Tasks 2011-2014.
Outline

Special Problems of Language Learners

Background: Corpora and Tasks

Heuristic and Data Driven Approaches

Annotation and Evaluation

Current and Future Trends
LEARNER ERRORS
Learner Errors: Cambridge Learner Corpus (CLC)

- Real word spelling
- Word order
- Run-on
- Agreement
- Pronoun
- Derivational morphology
- Verb formation
- Inflectional morphology
- Punctuation
- Determiner
- Preposition
- Content Word Choice

Bars representing the error percentages for each category.
Prepositions Presence and Choice: 13%

- Prepositions are problematic because they perform so many complex roles
  - Preposition choice in an adjunct is constrained by its object ("leave on Friday", "leave at noon")
  - Prepositions are used to mark the arguments of a predicate ("fond of beer.")
  - Phrasal Verbs ("give in to their demands.")
    - "give in" ⇔ "acquiesce, surrender"
Multiple prepositions can appear in the same context:

“When the plant is horizontal, the force of the gravity causes the sap to move __ the underside of the stem.”

<table>
<thead>
<tr>
<th>Choices</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>to</td>
<td>Writer</td>
</tr>
<tr>
<td>on</td>
<td>System</td>
</tr>
<tr>
<td>toward</td>
<td>Annotator 1</td>
</tr>
<tr>
<td>onto</td>
<td>Annotator 2</td>
</tr>
</tbody>
</table>
Determiner Presence and Choice: 12%

- English Article system: *a*, *an*, *the*
  - 7 levels of countability from *a car* to *an equipment*
  - Syntactic properties: *have a knowledge* vs *a knowledge of English*
- Discourse factors – previous mention
- Idioms: kick the/a bucket
- World Knowledge
  the moon (on earth)
Punctuation Conventions

- Apostrophe (1%):
  - Possessives
  - Contractions

- Comma (10%)
  - Missing after introductory clause
  - Fused sentences

AP Tweet: Dutch military plane carrying bodies from Malaysia Airlines Flight 17 crash lands in Eindhoven.

- Hyphenation (1%) when used adjectively
Verbal Morphology and Tense: 14%

- Over-regularization of irregular verbs
  - The women *weared/wore long dresses.

- Ill-formed tense, participle, infinitive, modal & auxiliary
  - I look forward to *see/seeing you.
  - People would *said/say
  - It can *do/be harmful.

- Can be dependent on discourse
  - I will clean my room yesterday
Derivational Morphology: 5%

- Confusion of adjectival, nominal, verbal, adverbial forms
  - I have already made the *arranged/arrangements.
  - There was a wonderful women volleyball match between Chinese team and *Cuba/Cuban team.
  - I *admiration/admire my teacher.
Pronoun Error: 4%

- Use of wrong case
  - *Him/He went to the store.

- Wrong gender
  - I met Jane and he showed me where to go.

- Vague pronoun reference
  - I’ll position the target, and when I nod my head, shoot at it.
Agreement Error: 4%

- These can be long distance
  - Three new texts which deal with this problem *has/have been written last year.

- Subject-verb agreement:
  - I *were/was in my house.

- Noun-number agreement
  - I am reading *these/this book.
  - Conversion always takes a lot of *efforts/effort.
Run-on Sentences: 4%

- Two independent clauses not connected by an appropriate punctuation or conjunction:
  - They deliver documents to them they provide fast service.
  - It is nearly half past five, we cannot reach town before dark.
Word Order (4%)

- Idiomatic
  - tried and true vs true and tried

- Ordering of adjectives & nominal compounds
  - A pop British band called “Spice Girl”.

- English word order: subject verb object (SVO)
  - Eat kids free (VSO)
Real Word Spelling Errors (2%)

- Homophones
  - there, their, they’re
  - to, too, two

- Near homophones
  - affect, effect
  - lose, loose
Content Word Choice: 20%

- Most common & least understood. Cover a wide range of errors & not fall into a pattern:
  - False friends: Eng rope / Sp ropa (clothes)
  - Collocation: strong / *powerful tea
    *strong / powerful computer
  - Confusion of similar looking or sounding words: Deliver the merchandise on a daily *base/basis.
  - ...

Influence of the Native Language

- L1 has no close equivalent construction – leading to difficulty in learning
  - Chinese and Russian have no equivalent of articles
- L1 has close equivalent construction – easier to learn.
  - German article system similar to English
- Two languages closely related – transfer problems where they differ
  - When Germans make article errors, likely a transfer problem
- Unrelated languages – no transfer but will make more errors due to difficulty of complex English structures
  - Chinese/Russians need to learn the article rules
- L1 works for and against a learner simultaneously
Goal of Grammatical Error Correction for Language Learners

- Grammatical error correction systems, like Microsoft Word, cover error types made by native speakers. They *rarely* identify article or preposition errors.

- Need systems that focus on those problems made by Language Learners: eg, articles, prepositions, verb formation, collocations, content word choice ...
Some Examples

http://www.tinyurl.com/kshecfw
BACKGROUND INFORMATION: CORPORA, EVALUATION & SHARED TASKS
Before discussing approaches, need some background:

- Identify non-proprietary corpora used in Grammatical Error Detection/Correction
- Review traditional NLP evaluation metrics
- 4 years of shared tasks/competitions
Corpora

- Until 2011, large learner corpora (1M or more words) were rare
  - And, except for Chinese Learners of English Corpus (CLEC), either proprietary or very expensive to license
- Since 2011, several have been made available
  - Enables cross-system evaluation
Differences in Corpora

- Corpora are not created equally ...
  - Different proficiency and L1
  - Different writing conditions (timed test vs. classroom assignment)
  - Different annotation standards
    - Annotators: native or non-native, experts or crowdsourced
    - Number of annotations per error – most have single annotation
  - Different annotation schemes
  - Different availability: licenses and fees
## Error-Annotated Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Description</th>
</tr>
</thead>
</table>
| **NUCLE** | - National University of Singapore Corpus of English  
- 1,450 essays by Singapore college students  
- Used in CoNLL shared tasks  
- Publically available |
| **FCE**  | - 1,244 essays from First Certificate in English exam (CLC subset)  
- Used in HOO 2012 task  
- Includes score, error annotation and demographics  
- Publically available |
| **HOO2011** | - Hand corrected papers from ACL Anthology  
- 38 conference papers  
- Publically available |
| **CLEC** | - Chinese Learners of English Corpus  
- 1M words  
- Five proficiency levels  
- Inexpensive |
Other Learner Corpora

- **TOEFL11**
  - ETS Corpus of Non-Native English
  - 12,100 essays (1,100 essays each for 11 different L1s)
  - Includes proficiency information
  - Available through LDC

- **ICLE**
  - International Corpus of Learner English
  - 3.7 M from over 16 different L1s
  - Partially error-annotated
  - Needs to be licensed.

- **Lang-8**
  - Language Learner Social Community Website
  - Nearly 200,000 Learner journal entries with community corrections
  - Need a script to extract data (Mizumoto et al., 2011)
I'm going to an afternoon mini-concert that's going to take place at the small stage in the shopping mall.

"the" may be correct here, but since we're not familiar with the shopping mall in question, "a" feels more natural.

My daughter in law is going to play the sax in it.

More of a suggestion than anything.

She's cool, active, and she's well into the last trimester of pregnancy, and proud.

This sentence sounds rather unnatural with the 'and' both here and at the end portion, so I removed this bit. Adding "of it" after "proud" will make that portion stronger and more specific, but may not be your intention.
TRADITIONAL NLP EVALUATION METRICS
Terminology

- **True Positive (TP) “hit”**
  - Flags *I am going for walk this afternoon.*

- **False Positive (FP)**
  - Flags *I am going for a walk this afternoon.*

- **True Negative (TN)**
  - Not flag *I am going for a walk this afternoon.*

- **False Negative (FN) “miss”**
  - Not flag *I am going for walk this afternoon.*
Traditional NLP Evaluation Metrics

\[
\text{Precision} = \frac{TPs}{TPs + FPs}
\]

\[
\text{Recall} = \frac{TPs}{TPs + FNs}
\]

\[
F\text{-score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

\[
\text{Accuracy} = \frac{TPs + TNs}{TPs + TNs + FPs + FNs}
\]
Precision, Recall and F-score are all used to evaluate shared tasks. However, they can be problematic for GEC evaluation and should be interpreted with caution – as discussed later.
Shared Tasks/Competitions

- Important for a field to progress
  - Helping Our Own (HOO): 2011 & 2012
  - Conference on Computational Natural Language Learning (CoNLL): 2013 & 2014
- Shared train and evaluation data sets
- Shared evaluation metrics

<table>
<thead>
<tr>
<th>Shared Task</th>
<th>Errors</th>
<th>Corpus</th>
<th># of Teams</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOO 2011</td>
<td>All</td>
<td>ACL Papers</td>
<td>6</td>
</tr>
<tr>
<td>HOO 2012</td>
<td>Preps &amp; Dets</td>
<td>FCE / (CLC)</td>
<td>14</td>
</tr>
<tr>
<td>CoNLL 2013</td>
<td>Preps, Dets, Nouns, Verbs</td>
<td>NUCLE</td>
<td>17</td>
</tr>
<tr>
<td>CoNLL 2014</td>
<td>All</td>
<td>NUCLE</td>
<td>12</td>
</tr>
</tbody>
</table>
Shared Task Evaluation Metrics

- **HOO 2011:** Three evaluations
  - Detection: Identify error
  - Recognition: Identify an error’s type and span
  - Correction: Provide at least one accurate rewrite
  - Precision, Recall & F-score calculated for each

- **HOO 2012:** Same as HOO 2011 plus
  - Participating teams could request changes in the annotation – adjudicated by organizers. Increased F-scores by almost 10%

- **CoNLL:** Same as HOO but different mapping
HOO Mapping

- HOO
  - Detection:
    - Any overlap with gold edit=TP
    - Output not overlap a gold edit=FP
    - No overlap with gold edit=FN
  - Recognition: edits must be exact
  - Correction: edits and labels must be exact
CoNLL Mapping

- MaxMatch (Dahlmeier & Ng, 2012): Edit Distance Measures (EDMs) to define errors over sequences of words
  - Maps, using EDMs, between system output and gold
  - Handles overlapping errors
  - Handles multiple sets of alternative corrections
Shared Tasks: Lessons Learned

- **Performance**
  - Despite 4 tasks, performance low: 20 to 40 F-score

- **Annotation Quality:**
  - Inconsistent
  - Systems penalized for valid corrections not annotated
  - Last 3 shared tasks allowed revisions to annotations by participants
    - The revisions increased F-score by almost 10%

- Need to deal with multiple interacting errors.
Different Approaches

A: Rule-Driven: No Context Needed

B: Rule-Driven: Local Context Needed

C: Parsing: Require syntactic structure, in sentence and beyond

D: Machine Learning methods

E: Whole Sentence Correction
A: No Context Needed: Simple as a Regular Expression

- Regular expressions for many verb errors:
  - Infinitive formation
    
    /to\s*(RB)\s*\s*VB\[DNGZ]\s*/ → /to\s*(RB)\s*\s*talk/

    to talking → to talk

  - Modal verb + have + past participle
    
    /MD\s*of\s*VBD\s*/ → /MD\s*have\s*VBD/

    would of liked → would have liked

- Word lists
  - Over-regularized morphology: I eated/ate an omelet.
B: Simple Statistically-based Approach: ALEK

- Unexpected combinations of POS tags:
  - Noun number /DT_a NNS/
    - I looked at a houses.

- Filtered by rules
  - Not trigger in the environment
    - /DT_a NNS NN/ -- a systems analyst

- Filtered by language model
  - Which is more likely, the original or the rewrite
1980’s: Before statistical parsers, modified parsers to recognize targeted errors

- Allow parse trees that violate constraints – increment counter. Best solution has smallest index.
- Add weights for specific violations
- Mal-rules: Write rules to detect specific errors
- Relax constraints on feature unification & use violations to produce feedback

None allow for analysis of arbitrarily ungrammatical text
Complex hand-crafted phrase structure rules that read off of a logical form.

```
write1 (+Pres +Pass +Perf +Proposition +Resultat +T1 +Space +Loc_sr)
 -Dsub_____X1
  -Dobj____text1 (+Pers3 +Sing +Mass)
   \Lcps_____three1 (+Quant +Plur +Num)
    \Attrib____new1 (+AO +Tme)
     \deal1 (+Inf +Pres +Proposition +IO)
      \Dsub____text1
       \with____problem1 (+Def +Prox1 +Pers3 +Sing +BndPrt +Count)
        \TmeAt_____year1 (+Def +Pers3 +Sing +BndPrt +Count +Tme +Tme_sr)
         \Attrib____last1 (+AO)
```

- Parse: Three new text which deal with the problem has been written.
- Pl quantifier and sg head noun. Suggest: text ➔ texts
- Parse: Three new texts which deal with the problem has been written.
- Detect subject-verb disagreement. Suggest: has ➔ have
- Parse: no error detected
C: Parsing/logical form for long distance dependencies

- Subject verb agreement
  - PP: The list of items is on the desk.
  - NP: Jack and Jill, who are late, are waiting on the corner.
  - RC: Barry the guy I met yesterday who has three kids lives in Brooklyn.

- Pronoun agreement
  - Nick and Marc were brothers and they live in Ireland.

- Run-on sentence/comma splice
  - They deliver documents to them they provide fast service.
D: Error types that Require Data-Driven Methods

- Articles (*a, an, the*): presence and choice
- Prepositions (10 – 27): presence and choice
- Auxiliary verbs (*be, do, have*): presence and choice
  - A fire will break out and it can do/*be harmful to people
  - A fire will break out and it can *do/be harmful to people.

- Gerund/Infinitive Confusion
  - On Saturday, I with my classmate went *eating/to eat.
  - Money is important in improving/*improve people's spirit.

Data-Driven Methods

Training Data

- Well-formed Text Only
- Artificial Errors
- Error-annotated Learner Data

Methods

- Classification
- Language Models
- Web-based
- Statistical Machine Translation

Well over 60+ papers!
Data-Driven Methods

Training Data
- Well-formed Text Only
- Artificial Errors
- Error-annotated Learner Data

Methods
- Classification
- Language Models
- Web-based
- Statistical Machine Translation
APPROACHES: CLASSIFICATION
Supervised classification requires:

- Machine learning classifier (MaxEnt, SVM, Average Perceptron, etc.)
- Data with labels for each training example

<table>
<thead>
<tr>
<th>Label</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>He will take our place <strong>in</strong> the line.</td>
</tr>
<tr>
<td>Error</td>
<td>He will take our place <strong>of</strong> the line.</td>
</tr>
</tbody>
</table>

Also need features!
Typical Features

- Parse:
  - He (PRP)
  - will (MD)
  - take (VB)
  - our (PRP$)
  - place (NN)
  - of (IN)
  - the (DT)
  - line (NN)

- POS:
  - subj
  - aux
  - dobj
  - poss
  - dobj
  - pobj
  - det

- Semantic:
  - WordNet
  - VerbNet
  - NER taggers
  - Semantic Role Labelers

- N-grams:
  - 1-gram: place, the
  - 2-gram: our-place, place-of, of-the, the-line
  - 3-gram: our-place-of, place-of-the, of-the-line

Source:
- Writer’s word(s) selection
- L1 of writer
- Genre of writing
Types of Training Data

1. Training on examples of correct usage only
2. Training on examples of correct usage and artificially generated errors
3. Training on examples of correct usage and real learner errors

Choice of training data largely determined by availability of data
1. Training on Correct Usage

- Prior to 2010, very few error-annotated corpora to get enough examples of errors for ML

- Solution: train on examples of correct usage only
  - [Han et al., 2006; Tetreault and Chodorow, 2008; Gamon et al., 2008; Felice and Pulman, 2009]

- Advantages: plenty of well-formed text available
  - Google n-gram corpus to build language models
  - Large corpora such as news, Wikipedia, etc. to derive features from

- Challenges:
  - Best to match genre of learner writing, so need lots of well-formed student essays
  - Does not exploit any information of when or how errors tend to appear
2. Artificial Errors

- Training only on examples of correct usage has performance limitations.

- Approximate learner writing by introducing artificial errors into a corpus of well-formed text.

- Training instances:
  - “Positive”: well-formed text
  - “Negative”: artificial errors

- Add a feature to capture transformation from erroneous choice to correct choice.

- Challenge: determining the best way to approximate the errors.
Rozovskaya and Roth (2010)

He will take our place in the line.

Method 1
Replace an article at random with various error rates

the \rightarrow a @ p(0.05)
the \rightarrow \text{null} @ p(0.05)

Method 2
Change distribution of articles so it is the same as in Learner text

Learner: (a, the, null) = (9.1, 22.9, 68.0)
Wiki: (a, the, null) = (9.6, 29.1, 61.4)

Method 3
Change distribution of articles so it is the same as in corrected Learner text

Learner: (a, the, null) = (9.5, 25.3, 65.2)
Wiki: (a, the, null) = (9.6, 29.1, 61.4)

Method 4
Change articles with learner error rate from annotated Learner text

the \rightarrow a @ p(0.14)
the \rightarrow \text{null} @ p(0.09)
Method 4 best; marginally more effective than training on well-formed text only (article errors)

- 10% error reduction in two cases
Artificial Errors

- Artificial error methodology was prominent in several shared task systems
- Felice et al. (EACL, 2014): expanded approach for other error types and other information (POS and sense)
- **GenERRate** (Foster and Andersen, 2009)
  - Tool for automatically inserting errors given a configuration file
3. Error-Annnotated Corpora

- Most common approach in shared tasks now that there are some labeled corpora available
- Use writer’s word choice as a feature
- Some key works:
  - Han et al. (2010): showed that having a large corpus of annotated essays significantly outperformed positive-examples-only training on prepositions
  - Dahlmeier & Ng (2011): showed that Alternating Optimization Techniques worked well with error-annotated data for prepositions
  - Most CoNLL 2014 shared task systems
Comparing Training Paradigms

- Izumi et al. (2003)
  - First to try all three training paradigms
  - Very little training data & focused on all errors→ results were poor
Comparing Training Paradigms

- Cahill et al. (2013)
  - Ten years later, try 3 paradigms again with multiple training and testing sets (Wikipedia Revisions, lang-8, NUCLE, FCE, news)
  - Focused on preposition errors only

- Trends:
  - Artificial errors derived from lang-8 proved best on 2 out of 3 test sets
  - Artificial error models can be competitive with real-error models, if enough training data generated
  - Training on Wikipedia revisions yields most consistent system across domains
APPROACHES:
WEB-BASED METHODS
Methods: Web-Based Methods

- Language learners will typically look at counts returned by search engine to figure out best word to use

- What happens when we use this simple methodology?
  - Select “target word” and search for alternatives
  - Select alternative with top web count
# Web-Based Methods

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Google Count</th>
<th>Bing Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>“fond of cats”</td>
<td>638,000</td>
<td>42,800</td>
</tr>
<tr>
<td>“fond for cats”</td>
<td>178</td>
<td>2</td>
</tr>
<tr>
<td>“fond by cats”</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>“fond to cats”</td>
<td>269</td>
<td>5</td>
</tr>
<tr>
<td>“fond with cats”</td>
<td>13,300</td>
<td>10</td>
</tr>
</tbody>
</table>
Methods:

- Prior work showed some value of the approach, but not over classification approaches.


Issues:

1. No POS tagging or lemma in search engines.
2. Search syntax is limited.
3. Constraints on number of queries per day.
4. Search counts are for pages not instances.
5. Search engines behave differently.
APPROACHES: LANGUAGE MODELS
Language Models

- Targeted Approach: can use LM scores over phrase or sentence for correction and detection

- Similar to Web-based approach though one has more control of the data

- Nearly half of the HOO2012 systems used LMs

He will take our place **in** the line.  

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| at | 0.1 |
| by | 0.2 |
| for | 0.1 |
| from | 0.0 |
| to | 0.1 |
| with | 0.1 |
Language Models

- Most commonly used in hybrid approaches:
  - As a “thresholder” for classification methods
  - Meta-learner: classification system weights decisions made by supervised classifier and LM (Gamon, 2010)
  - Rank whole sentence outputs from rule-based and SMT systems (Felice et al., 2014; Madnani et al., 2012)
APPROACHES:
STATISTICAL MACHINE TRANSLATION
Motivation

- Most work in correction targets specific error types such as prepositions and determiners
  - Large variety of grammatical errors in L2 writing
  - Errors often interact and overlap

- Can we use statistical machine translation (SMT) to do whole sentence error correction without requiring error detection?
  - Useful for feedback and content scoring

- Two types:
  - Noisy Channel
  - Round Trip Machine Translation
Noisy Channel Model

- View error correction as the process of *translating* from learner English to fluent English
  - Re-train MT system with examples of error phrases (or sentences) and their corrections
  - Dependent on having enough error-annotated data

Some examples:
- Brocket et al. (2006): use artificial errors to train SMT to correct mass noun errors
- Park & Levy (2011): use technique with FSTs
Round Trip Machine Translation

- Use pre-existing MT system to translate a sentence into another language and translate back into English
  - Thus does not use learner data
- Preliminary pilot studies with this method show some potential
Showed some promise with correcting French prepositions (Hermets and Desilets, 2009)

Showed some promise with whole sentence fluency correction (Madnani et al., 2012)
Other Issues

- Most prior work focused on specific errors (targeted approach)
- Targeted errors are easy to find when they are closed class or have a POS tag, but what happens in the case where they are missing?
  - “Some __ the people will be there.”
  - Can be difficult to detect
- Another issue: fixing awkward phrasings which span several words
Other Issues

- Most prior work focuses on prepositions and articles
  - Local features tend to be the most powerful

- Other errors are more complex:
  - Verb tense and aspect (Tajiri et al., 2012)
    - Require deeper understanding of sentence
    - Long range dependencies with verb forms in general
  - Collocations (Dahlemeier et al., 2012)
  - Content Words (Kochmar and Briscoe, 2014)
SYSTEM CASE STUDIES
1. Tetreault and Chodorow (2008)
   - Early example of an error correction methodology
   - Focused on preposition errors only
   - Trained on well-formed text

2. Rozovskaya et al. (CoNLL 2013)
   - Battery of classification approaches for 5 errors

3. Felice et al. (CoNLL 2014)
   - Combined SMT, rule-based and LM approach to handle all errors in 2014 Shared Task
1

TETREAULT & CHODOROW (2008): TARGETED ERROR APPROACH
Methodology

- Cast error detection task as a classification problem
- Given a model classifier and a context:
  - System outputs a probability distribution over 36 most frequent prepositions
  - Compare weight of system’s top preposition with writer’s preposition
- Error occurs when:
  - Writer’s preposition ≠ classifier’s prediction
  - And the difference in probabilities exceeds a threshold
Methodology

- Develop a training set of error-annotated learner essays (millions of examples?):
  - Too labor intensive to be practical

- Easy Alternative:
  - Train on millions of examples of proper usage

- Determining how “close to correct” writer’s preposition is
System Flow

Intermediate Outputs

- Essays
- Tokenized, POS, Chunk
- Preposition Features
- Errors Flagged

NLP Modules

- Pre-Processing
- Feature Extraction
- Classifier / Post-Processing

- 25 features built on lemma forms and POS tags
- Context consists of:
  - +/- two word window
  - Heads of the following NP and preceding VP and NP
<table>
<thead>
<tr>
<th>Feature</th>
<th>No. of Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PV</td>
<td>16,060</td>
<td>Prior verb</td>
</tr>
<tr>
<td>PN</td>
<td>23,307</td>
<td>Prior noun</td>
</tr>
<tr>
<td>FH</td>
<td>29,815</td>
<td>Headword of the following phrase</td>
</tr>
<tr>
<td>FP</td>
<td>57,680</td>
<td>Following phrase</td>
</tr>
<tr>
<td>TGLR</td>
<td>69,833</td>
<td>Middle trigram (pos + words)</td>
</tr>
<tr>
<td>TGL</td>
<td>83,658</td>
<td>Left trigram</td>
</tr>
<tr>
<td>TGR</td>
<td>77,460</td>
<td>Right trigram</td>
</tr>
<tr>
<td>BGL</td>
<td>30,103</td>
<td>Left bigram</td>
</tr>
</tbody>
</table>

He will take our place **in** the line
He will take our place in the line

<table>
<thead>
<tr>
<th>Feature</th>
<th>No. of Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PV</td>
<td>16,060</td>
<td>Prior verb</td>
</tr>
<tr>
<td>PN</td>
<td>23,307</td>
<td>Prior noun</td>
</tr>
<tr>
<td>FH</td>
<td>29,815</td>
<td>Headword of the following phrase</td>
</tr>
<tr>
<td>FP</td>
<td>57,680</td>
<td>Following phrase</td>
</tr>
<tr>
<td>TGLR</td>
<td>69,833</td>
<td>Middle trigram (pos + words)</td>
</tr>
<tr>
<td>TGL</td>
<td>83,658</td>
<td>Left trigram</td>
</tr>
<tr>
<td>TGR</td>
<td>77,460</td>
<td>Right trigram</td>
</tr>
<tr>
<td>BGL</td>
<td>30,103</td>
<td>Left bigram</td>
</tr>
</tbody>
</table>

He will **take** our **place** in the **line**
# Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>No. of Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PV</td>
<td>16,060</td>
<td>Prior verb</td>
</tr>
<tr>
<td>PN</td>
<td>23,307</td>
<td>Prior noun</td>
</tr>
<tr>
<td>FH</td>
<td>29,815</td>
<td>Headword of the following phrase</td>
</tr>
<tr>
<td>FP</td>
<td>57,680</td>
<td>Following phrase</td>
</tr>
<tr>
<td>TGLR</td>
<td>69,833</td>
<td><strong>Middle trigram (pos + words)</strong></td>
</tr>
<tr>
<td>TGL</td>
<td>83,658</td>
<td>Left trigram</td>
</tr>
<tr>
<td>TGR</td>
<td>77,460</td>
<td>Right trigram</td>
</tr>
<tr>
<td>BGL</td>
<td>30,103</td>
<td>Left bigram</td>
</tr>
</tbody>
</table>

He will take our **place** in the line.
Training Corpus

- Well-formed text → training only on positive examples
- 6.8 million training contexts total
  - 3.7 million sentences
- Two training sub-corpora:
  - MetaMetrics Lexile
    - 11th and 12th grade texts
    - 1.9M sentences
  - San Jose Mercury News
    - Newspaper Text
    - 1.8M sentences
Learner Testing Corpus

- Collection of randomly selected TOEFL essays by native speakers of Chinese, Japanese and Russian
- 8192 prepositions total (5585 sentences)
- Error annotation reliability between two human raters:
  - Agreement = 0.926
  - Kappa = 0.599
Full System

- Heuristic Rules that cover cases classifier misses
- Tradeoff recall for precision
- Pre-Filter: spelling, punctuation filtering
- Post-Filter: filter predictions made on antonyms, etc. and use manual rules for extraneous use errors
“He is fond with beer”
“My sister usually gets home by 3:00”
Performance

- Precision = 84%, Recall = 19%

- Typical System Errors:
  - Noisy context: other errors in vicinity
  - Sparse training data: not enough examples of certain constructions
ROZOVS KAYA ET AL. (2013): CONLL SHARED TASK SYSTEM
Overview

- **CoNLL 2013 Shared Task**
  - Correct *five* error types in NUCLE set
  - Art/Det, prepositions, nouns, verb form, verb agreement

- **System of five ML classifiers, one for each error type**
  - Aggregation of prior UIUC work
  - Finished 1<sup>st</sup> in *without-corrections* task ($F_{0.5} = 31.20$)
  - Finished 1<sup>st</sup> in *with-corrections* task ($F_{0.5} = 42.14$)
Basic Algorithm

- Preprocessing: POS and shallow parsing with UIUC tagger and chunker
- Methods for each error type:

<table>
<thead>
<tr>
<th>Error Type</th>
<th>ML</th>
<th>Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art/Det</td>
<td>Averaged Perceptron</td>
<td>NUCLE</td>
</tr>
<tr>
<td>Prepositions</td>
<td>Naïve Bayes</td>
<td>Google 1TB</td>
</tr>
<tr>
<td>Noun</td>
<td>Naïve Bayes</td>
<td>Google 1TB</td>
</tr>
<tr>
<td>Verb Form</td>
<td>Naïve Bayes</td>
<td>Google 1TB</td>
</tr>
<tr>
<td>Verb Agreement</td>
<td>Naïve Bayes</td>
<td>Google 1TB</td>
</tr>
</tbody>
</table>

- 4 lessons learned....
1. Learning Methods

- Experiments showed that Naïve Bayes with Google Web corpus regularly outperformed LMs for three error types.
2. Training Data

- Not always best to train on error-annotated data
- In the case of noun phrases, training on NUCLE was not as successful as using Google n-grams
3. Adaptation

- Provide error modules with knowledge of the error patterns of language learners
- Use adaptation (to change Naïve Bayes model priors) and artificial errors to improve performance for articles
4. Linguistic Knowledge

- For verb errors, determine which verbs are finite and non-finite
- Treat the two types differently
FELICE ET AL. (2014): CONLL SHARED TASK SYSTEM
Overview

- CoNLL 2014 Shared Task
  - Correct *all* error types in NUCLE set
- A system of multiple generation and ranking phases
  - Finished close 1st in *without-corrections* task ($F_{0.5} = 37.33$)
  - Finished close 2nd in *with-corrections* task ($F_{0.5} = 43.55$)
- Relies on rules, machine translation and LMs
Felice et al. (2014): Algorithm

Input: “Time changes, peoples change.”
Felice et al. (2014): Algorithm

Rule-Based System
- Rules extracted from CLC annotations
  - up to trigrams
- Morpho rules from dictionary
- High precision
Felice et al. (2014): Algorithm

Time changes, peoples change.
Time changes, people change.
Felice et al. (2014): Algorithm

LM Reranking
• 5-gram LM from Microsoft Web Services
Felice et al. (2014): Algorithm

SMT System
- Parallel Corpora (NUCLE, FCE, IELTS, Artificial Data)
- Trained with Moses and IRSTLM
Felice et al. (2014): Algorithm

Times change, and people change.
Felice et al. (2014): Algorithm

- RBS
- Generate Candidates
- LM
- SMT
- LM: Extract Corrections
- Type Filtering
- Apply Corrections

- Time ➔ Times
- changes ➔ change
- null ➔ and
- peoples ➔ people
Felice et al. (2014): Algorithm

- RBS
- Generate Candidates
- LM
- SMT
- LM
- Extract Corrections
- Type Filtering
- Apply Corrections

**Type Filtering**
- Heuristics from correction in NUCLE data
- Based on differences in word forms and POS
Felice et al. (2014): Algorithm

Times change, people change.
Summary of Felice et al. (2014)

- **Strengths:**
  - No need for distinct error modules
  - “One pass” approach
  - Handles interacting errors to an extent

- However, does rely partially on existence of enormous corpus of errors (proprietary CLC)
  - Makes it hard to generalize approach to other languages
Hands on Exercise

- Review Annotation Exercise
ANNO\'TATION & EVALUATION: TRIALS AND TRIBULATIONS
Overview

- Annotation
  - Annotation approaches: Comprehensive & Targeted
  - Multiple annotations per error
- Issues with Traditional NLP Evaluation Measure
- Rethinking Annotation and Evaluation with Crowdsourcing
Annotation Scheme 1: Comprehensive Approach

- Mark and correct all errors in the text

**Advantages:**
- Reliably estimates precision and recall

**Disadvantages:**
- Time consuming therefore expensive
- Error-prone as keep track of so many things at once
- Difficult to annotate adjacent and embedded errors:

  *In consion, for some reasons, museums, particuraly known travel place, get on many people.*
CLC Error Taxonomy

- About 80 error tags
  - 9 Word Classes: N=noun, J=adj, D=det, ...
  - 5 Modifications: wrong form (W), missing (M), needs replacing (R), unnecessary (U), wrongly derived (D)
  - Other error types include agreement, punctuation, spelling confusion, ...
  - UN=unnecessary noun
  - RJ=replace adjective
I arrived in time and the musical show started late so I was getting nervous because I dislike very much the impunctuality.

I arrived in time and the musical
<UN> delete “show”
started late so I was getting
<RJ> nervous → irritable
because I
<W> dislike very much → very much dislike
<UD> delete “the”
<DN> impunctuality → lateness
NUCLE Error Taxonomy

- 27 error tags
  - Verbs: tense, modal, missing, form
  - Subject-verb agreement
  - Article or Determiner
  - Nouns: number, possessive
  - Pronouns: form, reference
  - Word choice: wrong collocation/idiom/prep, wrong word form, wrong tone
  - Sentence: run-on/comma splice, dangling modifier, parallelism, fragment, ...
## Map CLC and NUCLE: Annotation Comparison

<table>
<thead>
<tr>
<th>Cambridge Learners Corpus</th>
<th>NUCLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing (MD), unnecessary (UD), or wrong (WD) determiner</td>
<td>ArtorDet</td>
</tr>
<tr>
<td>Missing (MP), unnecessary (UP), or wrong (WD) preposition</td>
<td>WordChoice</td>
</tr>
<tr>
<td>Missing (MV), unnecessary (UC), or wrong WV) verb, incorrect verb inflection</td>
<td>Verb tense (Vt), verb modal (Vm), missing verb (V0), Verb form (Vform)</td>
</tr>
</tbody>
</table>

- CLC: more descriptive but a very high cognitive load on annotators
- NUCLE: less descriptive but lower cognitive load on annotators
- Mapping between the two is a challenge
- Unlikely to get everyone to agree on a single tag set
Comprehensive Annotation Tool
(Rozovskaya and Roth, 2010)

Samuel Clemens is surely an author with a great gift for satirical and witty description of the life. In this book, partly based on his own reflections and experiences, he wants to present the local colour around Mississippi. The book deals with problems of the slavery system, which was an integral part of Clemens own childhood, even if he never considered it to be something unnatural, what is in fact the main problem of Huck as well.

Even if in his own way, he distinguishes between Heaven and Hell, sometimes he speaks to God and his final decision ‘‘ to go to Hell ‘’ is the fact of a great importance for him.

The whole book is inset with quotations from the Bible, which shows the attitude of the society to the religion. The second level of fantasy deals with different supernatural omens, widespread mostly among the black people, and is presented by the runaway slave.
Annotation Scheme 2: Targeted Approach

- If you want to develop a system/module that corrects a specific error type (e.g., preposition), comprehensive annotation is not required.

- Alternative: annotation on the target error type.

- Advantage of Focus: less cognitive load on annotator.

- For every error of that type:
  - Mark whether it is an error.
  - Insert alternative corrections.
  - Only need to mark errors in immediate context.
  - Indicate confidence in judgment.
### Example of Targeted Approach

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Status</th>
<th>Corrections</th>
</tr>
</thead>
<tbody>
<tr>
<td>The other thing I don't like going shopping <strong>in</strong> the weekend.</td>
<td>error</td>
<td>on, during</td>
</tr>
<tr>
<td>When I see some clothes <strong>in</strong> the window I like, I would go in and try them.</td>
<td>correct</td>
<td></td>
</tr>
<tr>
<td>When I see some clothes in the window I like, I would go <strong>in</strong> and try them.</td>
<td>correct</td>
<td></td>
</tr>
<tr>
<td>I am really appreciated it if you can tell <strong>to</strong> the people who work in Camp California, I choose the Golf and Photography.</td>
<td>error</td>
<td>null</td>
</tr>
<tr>
<td>I am really appreciated it if you can tell to the people who work <strong>in</strong> Camp California, I choose the Golf and Photography.</td>
<td>error</td>
<td>at</td>
</tr>
</tbody>
</table>
Implications of Using Multiple Annotators per document

- Advantages of multiple annotations per error:
  - Can report inter-annotator agreement – as well as system-annotator agreement
  - Identify error types that are difficult to annotate – annotator agreement can be low for some error types and high for others
  - Allows listing of more substitutions to improve evaluation

- Disadvantage: Annotation with two annotators per document is twice as expensive
Unexpected implications of using multiple annotators

- When using multiple annotations, serious issues with inter-annotator agreement become clear.
- In an experiment by Tetreault & Chodorow (2008), depending on the annotator, results differed by 10% precision and 5% recall.
How to make annotation more efficient and more accurate?

- We will come back to this – tying in both annotation and evaluation
- First an overview of issues with evaluation
ISSUES WITH EVALUATION
Issues with Evaluation

1. Mapping from system output to gold standard
2. Cautions about traditional metrics
3. MaxMatch CoNLL mapping scheme
Issue #1

Some of people like cats.

Some of the people like cats.

Some people like cats.
Issue #1: Mapping from Writer’s Errors to Gold Standard

- More than one way to label and repair an error
  - Book inspired me
    - Article error: A book inspired me.
    - Noun-number error: Books inspired me.
    - Both: The book inspired me.

- More than one way to repair an error
  - It can do harmful:
    - It can do harm
    - It can be harmful
  - I sat on the sunshine. → Can rewrite with in or under
## Manual Verification against CLC

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrects a CLC error</td>
<td>33%</td>
</tr>
<tr>
<td>Corrects an error that was not annotated as being an error in CLC</td>
<td>12%</td>
</tr>
<tr>
<td>Corrects a CLC error, but uses an alternative, but acceptable, correction</td>
<td>4%</td>
</tr>
<tr>
<td>Original and suggested correction are equally good</td>
<td>10%</td>
</tr>
<tr>
<td>Error correctly detected, but the correction is wrong</td>
<td>9%</td>
</tr>
<tr>
<td>Identifies an error site, but the actual error is not a preposition error</td>
<td>19%</td>
</tr>
<tr>
<td>Introduces an error</td>
<td>15%</td>
</tr>
</tbody>
</table>
Verification results

- Before manual verification, accuracy against annotations is 33%
- After manual verification, only 14% of the corrections are False Positives
- HOO and CoNLL evaluations evolved to mitigate these effects
- Progress has been made – but we’re not there yet
Issue #2: Traditional Evaluation

- **Accuracy** can be misleading
  - Learner error rates are low across entire corpus
  - Large TN values dominate calculation

- Example
  - IF preposition errors occur 10% of time in a learner corpus
  - THEN a baseline system that always treats prepositions as correct has 90% accuracy
Issue #2: Traditional Evaluation

- **Recall** not account for chance
- Prevalence (skewness of data)
  - A system that performs at chance will show increased recall when there is an increase in the proportion of cases *annotated as errors* (Powers, 2012)
  - Can’t compare systems that use different corpora
- **Bias**
  - A system that performs at chance will show increased recall when there is an increase in the proportion of cases *flagged as errors* – even when they are FPs
  - Can’t compare systems that generate flags at different rates
Cohen’s Kappa – Account for chance

- Subtract proportion expected by chance \((P_e)\) from the observed agreement \((P_o)\)

\[
K = \frac{P_o - P_e}{1 - P_e}
\]

- Result is a fraction between 0 and 1.0
  - 0 = no agreement & 1.0 = perfect agreement
  - 0.20 – 0.40 = slight agreement
  - 0.40 – 0.60 = moderate agreement
  - 0.60 and above = substantial agreement
Kappa values depend largely on how TNs are counted

- Calculations rely heavily on the number True Negatives, which can be computed in many ways.
- How many TNs for omitted articles where the system suggests inserting *a* before *walk*?
  
  I am going for *walk* this *afternoon*.
  
  - 6 if every word is a possible site
  - 3 if every NP is a possible site
  - 2 if pronouns are not a possible site
  - 1 if neither pronouns nor determiners are possible sites
Issue #3: CoNLL MaxMatch

- Precision, Recall & F-score computed using the MaxMatch algorithm (Dahlmeier & Ng, 2012)

- Drawbacks:
  - Focus on comparing strings rather than source & type of error – it is harder to provide feedback to learners
  - Chodorow et al (2012): No way to derive TNs and thus to compute Kappa statistic
Evaluation Metrics: Proposed Guidelines

- With so many metrics, and others on the way, use these guidelines (Chodorow et al 2012):
  - Report raw numbers of True Positives, False Positives, False Negatives, True Negatives
  - Be clear about how you calculate True Negatives
- Report statistical significance
RETHINKING ANNOTATION AND EVALUATION WITH CROWDSOURCING
Crowdsourcing

- Advantages: fast & cheap source of untrained annotators
- Has been used successfully in many NLP tasks:
  - Word Sense Disambiguation, Sentiment Analysis, etc.
- Can be used to address several deficiencies in annotation and thus evaluation:
  - Multiple raters: can be used to better create gold standard(s)
  - Time and therefore cost
Preposition Error Annotation (Tetreault et al., 2010)

- Rate the preposition!

  He feels bad about him and will be living rugged and lonely life.
  - Preposition is correct
  - Preposition is incorrect
  - Preposition is too hard to judge given the words surrounding it

- Results
  - 3 annotators $K = 0.61$
  - 13 Turker/annotator $K = 0.61$
Quality Control Experiment
(Tetreault et al., 2013)

- Replicate and extend earlier experiments using Crowdflower that screens out unreliable Turkers
- Result: even fewer Turkers required (though comes at a higher price)

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Amazon Mechanical Turk</th>
<th>CrowdFlower</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepositions</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Determiners</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Collocations</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
Rethinking Annotation & Evaluation

- Prior evaluations rest on the assumption that all prepositions are of equal difficulty
- However, some contexts are easier to judge than others:

  **Easy**
  - “It depends *of* the price of the car”
  - “The only key *of* success is hard work.”

  **Hard**
  - “Everybody feels curiosity *with* that kind of thing.”
  - “I am impressed that I had a 100 score *in* the test of history.”
  - “Approximately 1 million people visited the museum in Argentina *in* this year.”
Difficulty of cases can skew performance and system comparison

If System X performs at 80% on corpus A, and System Y performs at 80% on corpus B →

- ...Y is probably the better system
- But need difficulty ratings to determine this
Rethinking Annotation & Evaluation
(Madnani et al., 2011)

- Group errors into “difficulty bins” based on AMT agreement
  - 90% bin: 90% of the Turkers agree on the rating for an error (strong agreement)
  - 50% bin: Turkers are split on the rating for an error (low agreement)
- Run system on each bin separately and report results
- Gives more weight to cases with high human agreement
Summary

- Clearly more research needs to be done with
  - Different error types
  - Different designs/interfaces
- BUT this is will likely be a fruitful avenue for future research. Annotating in a fraction of the time at a fraction of the cost.
CURRENT & FUTURE DIRECTIONS
Current State of Affairs

- Shared Resources
  - Shared Tasks

- Workshops
  - Lots of papers
  - Two M&C Books

But: performance still quite low relative to other NLP tasks! Where do we go from here?
What is the future of GEC?

- A high performance system which can detect and classify grammatical errors by a language learner.
What is the future of GEC?

Provide useful feedback to learner

Track learner *over time* and model language development

Take into account L1, user context, etc.

Integrate with persistent spoken dialogue tutor
What is the future of GEC?

- A system which can automatically transform one noisy sentence to a fluent sentence...without a change in meaning.

Having discuss all this I must say that I must rather prefer to be a leader than just a member.

After discussing all this I must say that I’d prefer to be a leader than a follower.
What is the future of GEC?

System need not simply be a text to text transformation, could also take into account:

- Other sentences in document
- Context of document (writer’s intention)
  - Register
  - Who the document is for
- Prior sentences writer has produced (personalization)
SHORTER TERM DIRECTIONS
Annotation for Evaluation

- Despite development of new corpora, annotation and evaluation best practices still an open issue
- How to efficiently and cheaply collect high quality judgments?
- How to collect multiple judgments on a sentence?
- How to use multiple judgments for evaluation?
  - Borrow from MT evaluation field
- Best metrics to use? [Chodorow et al., 2012]
Multilingual GEC

- GEC for other languages hampered by:
  - Lack of good NLP tools (taggers, parsers, etc.)
  - Lack of large corpora (even of well formed text)
  - Lack of evaluation data
- Need to explore other techniques: web-scraping, Wikipedia Revisions, lang-8 hold promise, though might not be large enough
  - Israel et al. (2013) – Korean error correction
  - CLP Shared Task on Chinese as a Foreign Language
  - EMNLP Shared Task on Automatic Arabic Error Correction
Other Error Types

- Most work has focused on prepositions and articles
- Still other error types: verbs, collocations, word choice, punctuation, etc. which have very little research behind them

<table>
<thead>
<tr>
<th>Error Type</th>
<th># of Errors</th>
<th>Best Team</th>
<th>F-score</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArtorDet</td>
<td>690</td>
<td>UIUC</td>
<td>33.40</td>
<td>Avg Perceptron</td>
</tr>
<tr>
<td>Prep</td>
<td>312</td>
<td>NARA</td>
<td>17.53</td>
<td>MaxEnt</td>
</tr>
<tr>
<td>NN</td>
<td>396</td>
<td>UIUC</td>
<td>44.25</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>Vform/SVA</td>
<td>246</td>
<td>UIUC</td>
<td>24.51</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>Overall</td>
<td>1644</td>
<td>UIUC</td>
<td>31.20</td>
<td>Collection</td>
</tr>
</tbody>
</table>

CoNLL 2013 Shared Task Results
Most work treats error correction as a process sitting on an NLP pipeline of POS-tagging and parsing.

However, changing / adding / deleting words can alter POS tags and parse structure.

Do error correction and POS tagging/parsing as joint model (Sakaguchi et al., COLING 2012)
Joint Models for Error Correction

- Most work treats error correction as a collection of individual, usually *independent* modules.
- Addressing one error may have a ripple effect on another error:
  - Tense changes
  - “They believe that *such situation* must be avoided.”
- Some recent work:
  - Dahlmeier & Ng (2013): beam search decoding
  - Rozovskaya et al. (2014): joint inference
As we saw earlier, some preliminary work which incorporates L1

- Hermet and Desilets (2009)
- Tetreault and Chodorow (2009)
- Rozovskaya and Roth (2011)

Line of research in its infancy due to data scarcity
Unsupervised Methods

- Nearly all current work uses some form of supervision
- Lots of unlabeled learner data available:
  - Learner websites and forums
  - Lang-8
  - TOEFL11 corpus
- How can these sources be leveraged?
  - Levy and Park (2011)
Direct Application of GEC

- Bulk of work has focused on “test tube” evaluations of GEC
- But how do GEC systems impact student learning in the short term and long term?
- NLP field should start connecting with Second Language Learning and education researchers
  - Have students use GEC system in the classroom (Criterion)
  - Incorporate GEC into dialogue tutoring system
Applications of GEC

- Automated Essay Scoring
  - Attali and Burstein (2006)

- Native Language Identification
  - Koppel et al. (2005), Tetreault et al. (2012)

- MT Quality Estimation
  - Bojar et al. (2013), Callison-Burch et al. (2012)

- Noisy data processing
  - Social Media / normalization
  - MT post-processing
  - Assistive Tech: GEC of automatic closed captions
This tutorial:
- Provided a history of GEC
- Described popular methodologies for correcting language learner errors
- Described issues with annotation and evaluation

Grammatical Error Correction one of the oldest fields and applications of NLP

Still much work to be done as performance is still low!
Acknowledgments

- Martin Chodorow
- Michael Gamon
- Mariano Felice and the Cambridge Team
- Nitin Madnani
- Mohammad Sadegh Rasooli
- Alla Rozovskaya
Resources

- **HOO Shared Tasks**
- **CoNLL 2013 Shared Task**
- **CoNLL 2014 Shared Task**
- **BEA Workshop Series**

New 2014 Version