The Ups and Downs of Preposition Error Detection in ESL Writing

Joel Tetreault [Educational Testing Service]

What does ETS do?

- Standardized Assessment
  - GRE
  - TOEFL
  - TOEIC
  - SAT
  - Others

EVIL

- Educational Tools
  - Criterion, Text Adaptor
- Educational Policy

A Brief History of ETS

1930s-1940s | 1950-1980s | 1990s - 2000s | Present

- 1930s: to get into university, one had to be wealthy or attend top prep schools
- Henry Chauncey believed college admission should be based on achievement, intelligence
- With other Harvard faculty, created standardized tests for military and schools
- ETS created in 1947 in Princeton, NJ

A Brief History of ETS

1930s-1940s | 1950-1980s | 1990s - 2000s | Present

- ETS grows into the largest assessment institution
- SAT and GRE are biggest tests, with millions of students over 180 countries taking them each year
- Make move from multiple choice to more natural questions (essays)

NLP Meets Assessment

1930s-1940s | 1950-1980s | 1990s - 2000s | Present

- Revenue
  - Cost Savings for Large-Scale Assessments
  - Market for Practice Instruction & Assessments
- Classroom Teacher Support for Writing
  - More practice writing possible
  - Individual and classroom performance assessment
  - Electronic writing portfolios

NLP Meets Assessment

1930s-1940s → 1950-1980s → 1990s - 2000s → Present

- E-rater / *Criterion*℠ (essay scoring)
- C-rater (short answer content scoring)
- Speech Rater (speech scoring)
- *Text Adaptor* (teacher assistance tools)
- Plagiarism Detection

E-rater

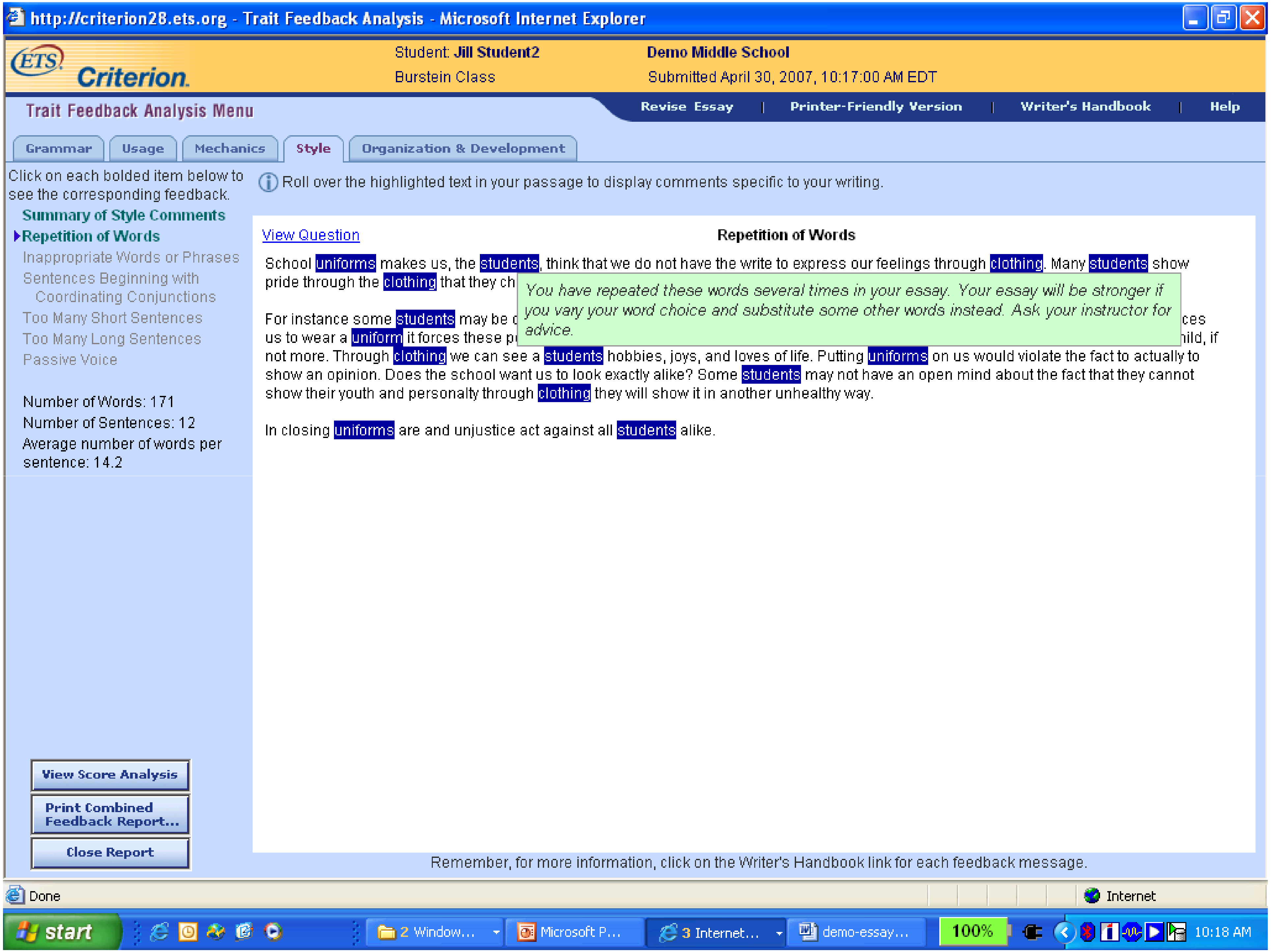
- First deployed in 1999 for GMAT Writing Assessment
- System Performance:
  - *E-rater*/Human agreement: 50% exact, 90% exact (+1 adjacent)
  - Comparable to two humans
- Massive collection of 50+ weighted features organized into 5 high level features
- Combined using stepwise linear regression

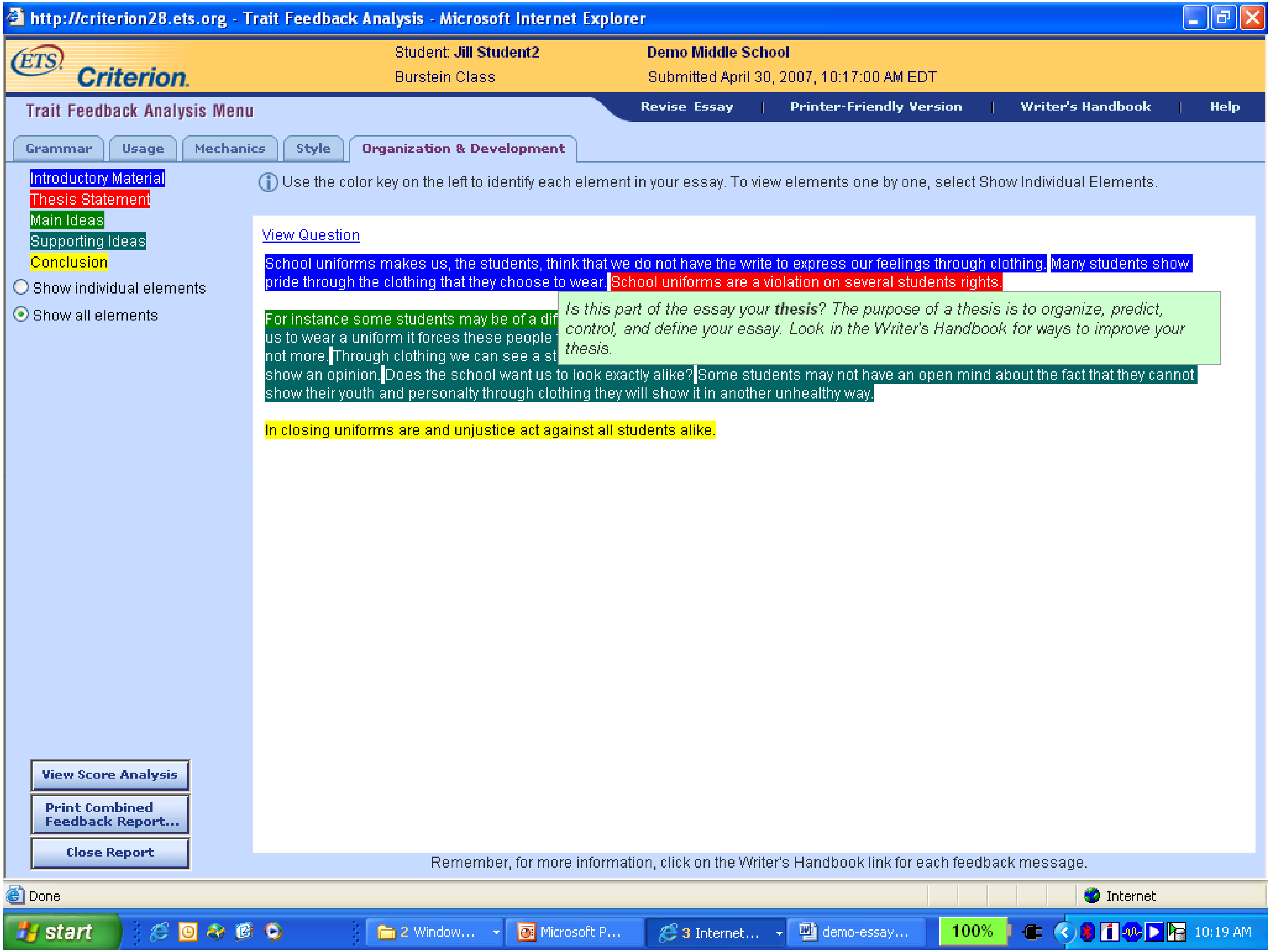
E-rater Features

Grammar	• Sentence fragments, garbled words • Pronoun, possessive errors
Usage	• Wrong word form, double negative • Incorrect article/preposition
Mechanics	• Spelling • Punctuation
Style	• Sentence length, word repetition • Passives
Organization	• Discourse sequences • RST & Syntactic structures

Criterion

- E-rater as classroom instruction/feedback tool
- Used in 3200+ schools
- Over 3M submissions since 2001
- Over 1M student registrations
- International Use:
  - Canada, Mexico, India, Puerto Rico, Egypt, Nepal, Taiwan, Hong Kong, Japan, Thailand, Vietnam, Brazil, UK, Greece, Turkey


http://criterion28.ets.org - Trait Feedback Analysis - Microsoft Internet Explorer
ETS Criterion
Student: Jill Student2
Burstein Class
Demo Middle School
Submitted April 30, 2007, 10:17:00 AM EDT
Trait Feedback Analysis Menu
Revise Essay | Printer-Friendly Version | Writer's Handbook | Help
Grammar
Usage
Mechanics
Style
Organization & Development
Click on each bolded item below to see the corresponding feedback.
Summary of Style Comments
Repetition of Words
Inappropriate Words or Phrases
Sentences Beginning with Coordinating Conjunctions
Too Many Short Sentences
Too Many Long Sentences
Passive Voice
Number of Words: 171
Number of Sentences: 12
Average number of words per sentence: 14.2
View Score Analysis
Print Combined Feedback Report...
Close Report
Roll over the highlighted text in your passage to display comments specific to your writing.
View Question
Repetition of Words
School uniforms makes us, the students, think that we do not have the write to express our feelings through clothing. Many students show pride through the clothing that they ch
You have repeated these words several times in your essay. Your essay will be stronger if you vary your word choice and substitute some other words instead. Ask your instructor for advice.
For instance some students may be c
ces
us to wear a uniform it forces these p
hild, if
not more. Through clothing we can see a students hobbies, joys, and loves of life. Putting uniforms on us would violate the fact to actually to show an opinion. Does the school want us to look exactly alike? Some students may not have an open mind about the fact that they cannot show their youth and personalty through clothing they will show it in another unhealthy way.
In closing uniforms are and unjustice act against all students alike.
Remember, for more information, click on the Writer's Handbook link for each feedback message.
Done
Internet
start
2 Window...
Microsoft P...
3 Internet...
demo-essay...
100%
10:18 AM


http://criterion28.ets.org - Trait Feedback Analysis - Microsoft Internet Explorer
ETS Criterion
Student: Jill Student2
Burstein Class
Demo Middle School
Submitted April 30, 2007, 10:17:00 AM EDT
Trait Feedback Analysis Menu
Revise Essay | Printer-Friendly Version | Writer's Handbook | Help
Grammar
Usage
Mechanics
Style
Organization & Development
Introductory Material
Thesis Statement
Main Ideas
Supporting Ideas
Conclusion
Show individual elements
Show all elements
Use the color key on the left to identify each element in your essay. To view elements one by one, select Show Individual Elements.
View Question
School uniforms makes us, the students, think that we do not have the write to express our feelings through clothing. Many students show pride through the clothing that they choose to wear. School uniforms are a violation on several students rights.
For instance some students may be of a dif
us to wear a uniform it forces these people
not more. Through clothing we can see a st
show an opinion. Does the school want us to look exactly alike? Some students may not have an open mind about the fact that they cannot show their youth and personalty through clothing they will show it in another unhealthy way.
Is this part of the essay your thesis? The purpose of a thesis is to organize, predict, control, and define your essay. Look in the Writer's Handbook for ways to improve your thesis.
In closing uniforms are and unjustice act against all students alike.
View Score Analysis
Print Combined Feedback Report...
Close Report
Remember, for more information, click on the Writer's Handbook link for each feedback message.
Done
Internet
start
2 Window...
Microsoft P...
3 Internet...
demo-essay...
100%
10:19 AM

What's Next for ETS?

- Assessment/tools for learners of English as a Second Language (ESL)
  - 300 million ESL learners in China alone
  - 10% of US students learn English as a second language
  - Teachers now burdened with teaching classes with wildly varying levels of English fluency

What's Next for ETS?

1930s-1940s → 1950-1980s → 1990s - 2000s → Present

- Increasing need for tools for instruction in English as a Second Language (ESL)
- Other Interest:
  - Microsoft Research (ESL Assistant)
  - Publishing Companies (Oxford, Cambridge)
  - Universities
  - Rosetta Stone

Objective

- Long Term Goal: develop NLP tools to automatically provide feedback to ESL learners about grammatical errors

- Preposition Error Detection
  - Selection Error ("They arrived *to* the town.")
  - Extraneous Use ("They came *to* outside.")
  - Omitted ("He is fond this book.")

Preposition Error Detection

- Present a combined ML and rule-based approach:
  - State of the art performance in native & ESL texts
- Similar methodology used in:
  - Microsoft's ESL Assistant [Gamon et al., '08]
  - [De Felice et al., '08]
- This work is included in ETS's *Criterion*SM Online Writing Service and *E-Rater* (GRE, TOEFL)

Outline

1. Motivation
2. Approach
   - Methodology
   - Feature Selection
3. Evaluation on Native Text (Prep. Selection)
4. Evaluation on ESL Text
5. Future Directions

Motivation

- Preposition usage is one of the most difficult aspects of English for non-native speakers
  - [Dalgish ’85] – 18% of sentences from ESL essays contain a preposition error
  - Our data: 8-10% of all prepositions in TOEFL essays are used incorrectly

Why are prepositions hard to master?

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

Why are prepositions hard to master?

- Multiple prepositions can appear in the same context:

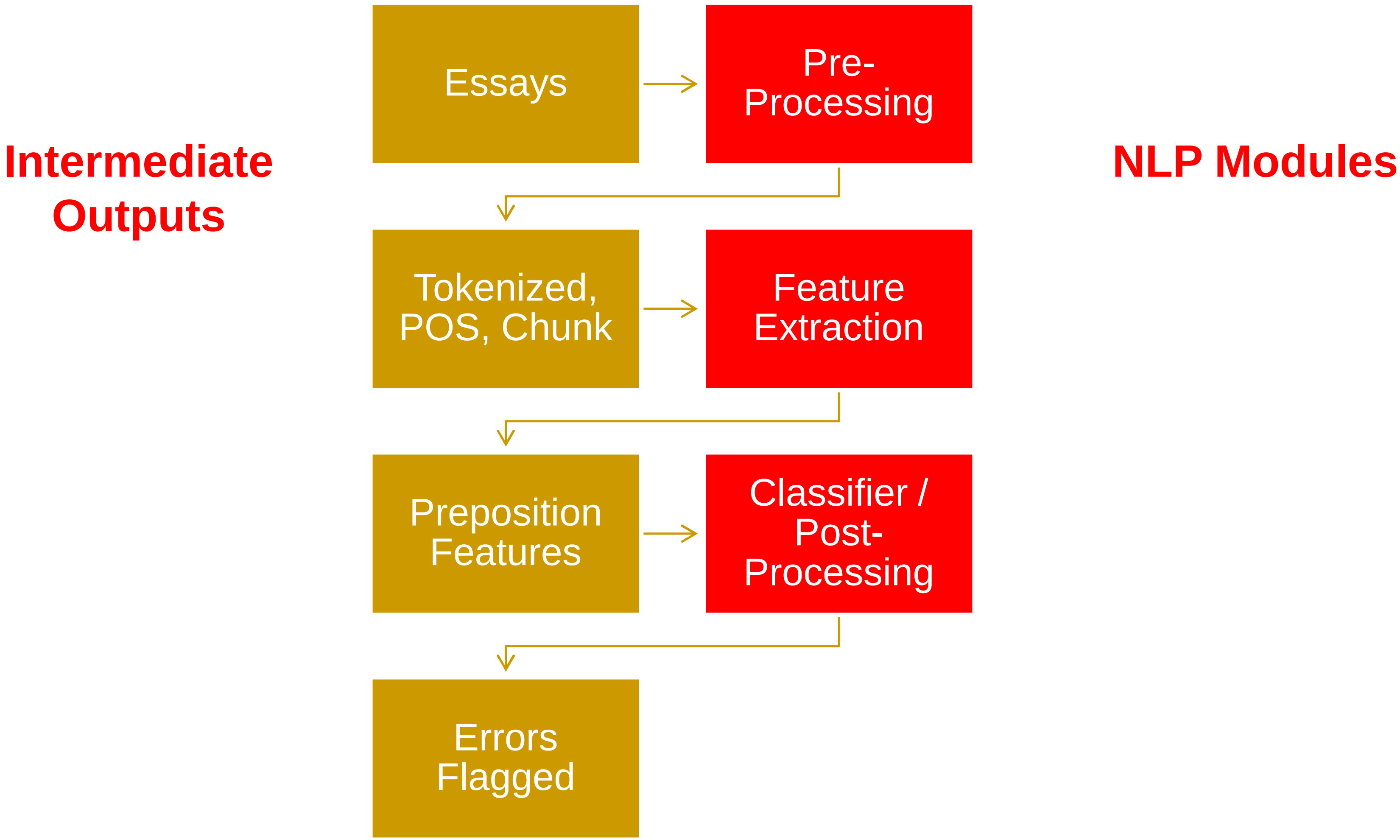
"When the plant is horizontal, the force of the gravity causes the sap to move __ the underside of the stem."

Choices	Source
• to • on • toward • onto	• Writer • System • Rater 1 • Rater 2

NLP & Preposition Error Detection

1. Methodology for Preposition Error Detection
   - [Tetreault & Chodorow, COLING ’08]
   - [Chodorow, Tetreault & Han, SIGSEM-PREP ‘07]
   - [Tetreault & Chodorow, WAC ’09]
2. Experiments in Human Annotation
   - Implications for system evaluation
   - [Tetreault & Chodorow, HJCL ‘08]

System Flow


Intermediate Outputs
NLP Modules
Essays
Pre-Processing
Tokenized, POS, Chunk
Feature Extraction
Preposition Features
Classifier / Post-Processing
Errors Flagged

Methodology

- Cast error detection task as a classification problem
- Given a model classifier and a context:
  - System outputs a probability distribution over 34 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 ESL essays (millions of examples?):
  - Too labor intensive to be practical
- Alternative:
  - Train on millions of examples of proper usage
- Determining how “close to correct” writer’s preposition is

Feature Selection

- Prepositions are influenced by:
  - Words in the local context, and how they interact with each other (lexical)
  - Syntactic structure of context
  - Semantic interpretation

Feature Extraction

- Corpus Processing:
  - POS tagged (Maxent tagger [Ratnaparkhi '98])
  - Heuristic Chunker
  - Parse Trees?
    - "In consion, for some reasons, museums, particuraly known travel place, get on many people."
- Feature Extraction
  - Context consists of:
    - +/- two word window
    - Heads of the following NP and preceding VP and NP
  - 25 features consisting of sequences of lemma forms and POS tags

Features

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

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

Features

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

He will $\boxed{\text{take}}$ our $\boxed{\text{place}}$ **in** the $\boxed{\text{line}}$

PV **PN** **FH**

Features

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

He will take our $\boxed{\text{place } \textbf{in} \text{ the}}$ line.

TGLR

Combination Features

- MaxEnt does not model the interactions between features
- Build “combination” features of the head nouns and commanding verbs
  - PV, PN, FH
- 3 types: word, tag, word+tag
  - Each type has four possible combinations
  - Maximum of 12 features

Combination Features

Class	Components	+Combo:word
p-N	FH	line
N-*p*-N	PN-FH	place-line
V-*p*-N	PV-PN	take-line
V-N-*p*-N	PV-PN-FH	take-place-line

“He will take our place in the line.”

Preposition Selection Evaluation

- Test models on well-formed native text
- Metric: accuracy
  - Compare system's output to writer's
  - Has the potential to underestimate performance by as much as 7% [HJCL '08]
- Two Evaluation Corpora:

WSJ	Encarta-Reuters
test=106k events train=4.4M NANTC events	test=1.4M events train=3.2M events Used in [Gamon+ '08]

Preposition Selection Evaluation

Model	WSJ	Enc-Reu*
Baseline (of)*	26.7%	27.2%
Lexical	70.8%	76.5%
+Combo	71.8%	77.4%
+Google	71.6%	76.9%
+Both	72.4%	77.7%
+Combo +Extra Data	74.1%	79.0%

* [Gamon et al., ’08] perform at 64% accuracy on 12 prep’s

Evaluation on Non-Native Texts

- Error Annotation
  - Most previous work used only one rater
  - Is one rater reliable? [HJCL '08]
  - Sampling Approach for efficient annotation
- Performance Thresholding
  - How to balance precision and recall?
  - May not want to optimize a system using F-score
- ESL Corpora
  - Factors such as L1 and grade level greatly influence performance
  - Makes cross-system evaluation difficult

Training Corpus for ESL Texts

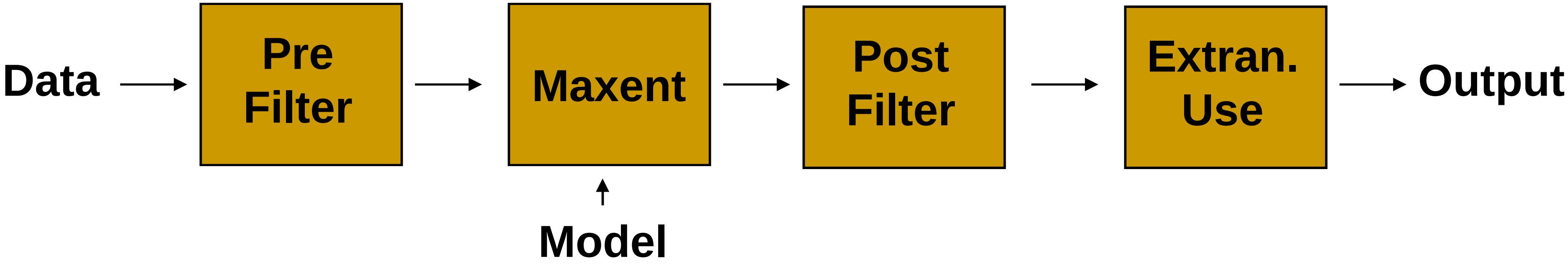
- Well-formed text → training only on positive examples
- 6.8 million training contexts total
  - 3.7 million sentences
- Two training sub-corpora:

MetaMetrics Lexile	San Jose Mercury News
❑ 11th and 12th grade texts ❑ 1.9M sentences	❑ Newspaper Text ❑ 1.8M sentences

ESL Testing Corpus

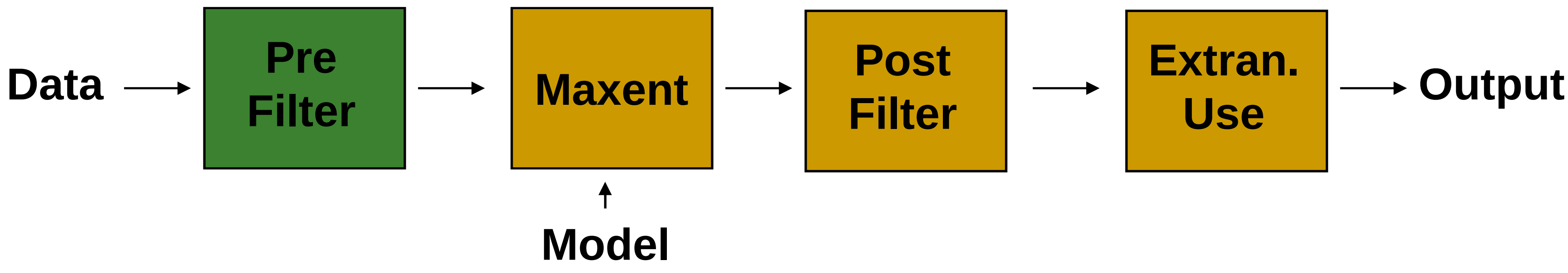
- Collection of randomly selected TOEFL essays by native speakers of Chinese, Japanese and Russian
- 8192 prepositions total (5585 sentences)
- Error annotation reliability between two human raters:
  - Agreement = 0.926
  - Kappa = 0.599

Expanded Classifier



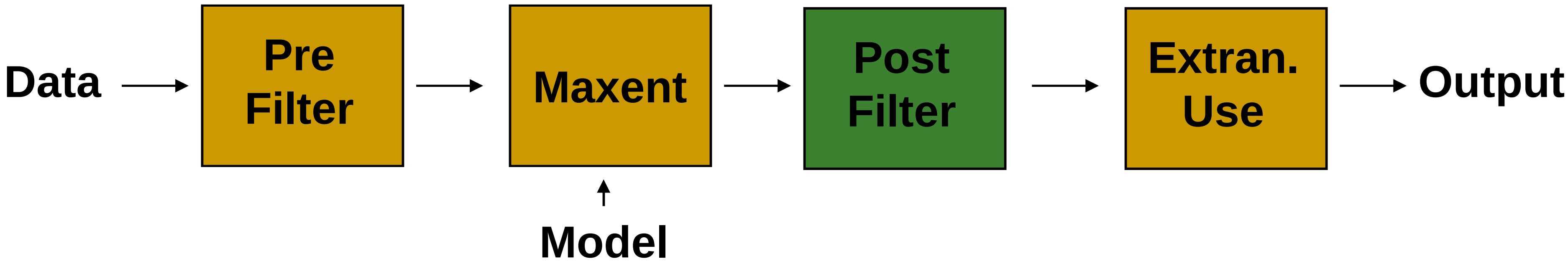
- Pre-Processing Filter
- Maxent Classifier (uses model from training)
- Post-Processing Filter
- Extraneous Use Classifier (PC)

Pre-Processing Filter



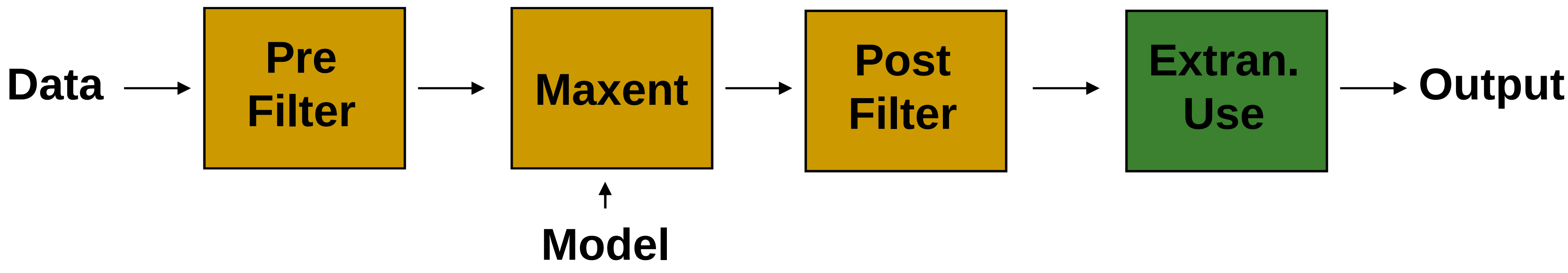
- Spelling Errors
  - Blocked classifier from considering preposition contexts with spelling errors in them
- Punctuation Errors
  - TOEFL essays have many omitted punctuation marks, which affects feature extraction
- Tradeoff recall for precision

Post-Processing Filter



- Antonyms
  - Classifier confused prepositions with opposite meanings (with/without, from/to)
  - Resolution dependent on intention of writer
- Benefactives
  - Adjunct vs. argument confusion
  - Use WordNet to block classifier from marking benefactives as errors

Prohibited Context Filter



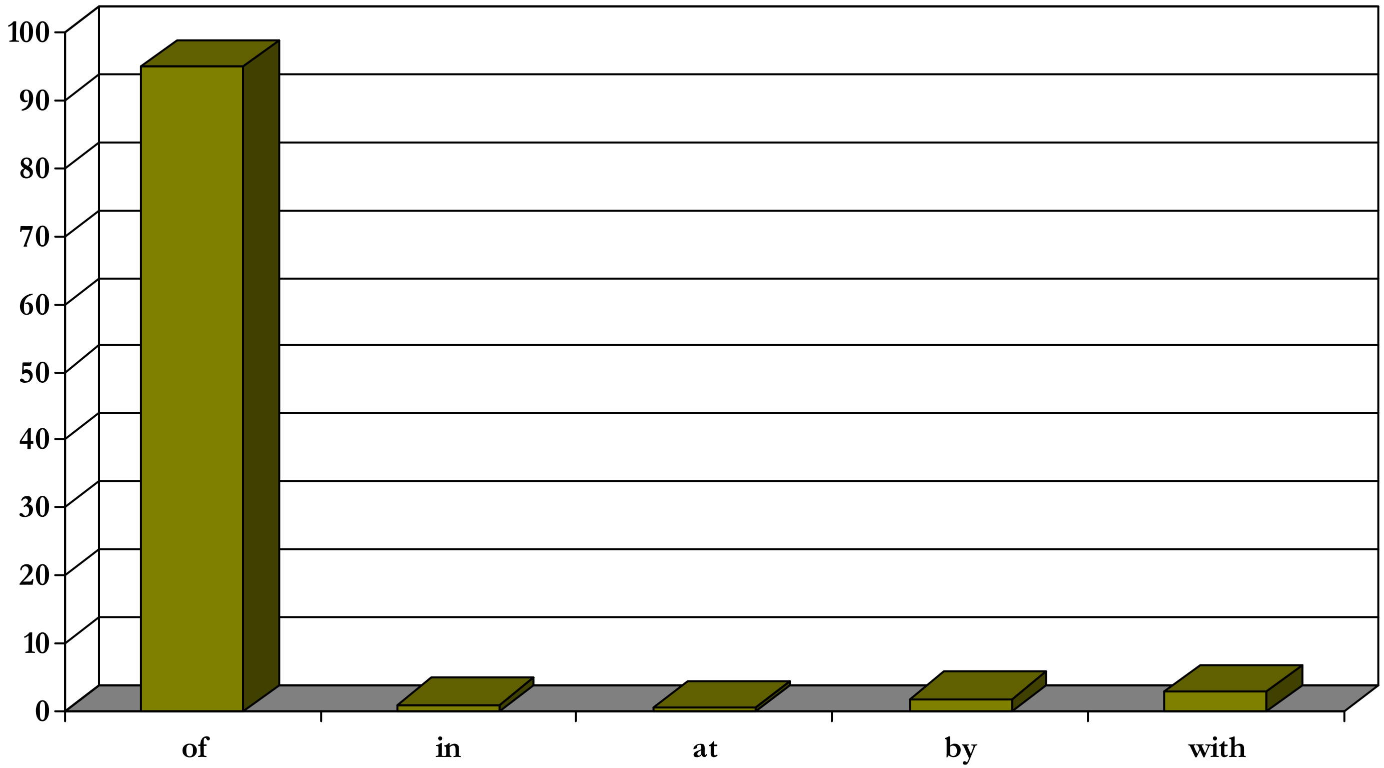
- Account for 142 of 600 errors in test set
- Two filters:
  - Plural Quantifier Constructions ("some *of* people")
  - Repeated Prep's ("can find friends *with with*")
- Filters cover 25% of 142 errors

Thresholding Classifier's Output

- Thresholds allow the system to skip cases where the top-ranked preposition and what the student wrote differ by less than a pre-specified amount

Thresholds

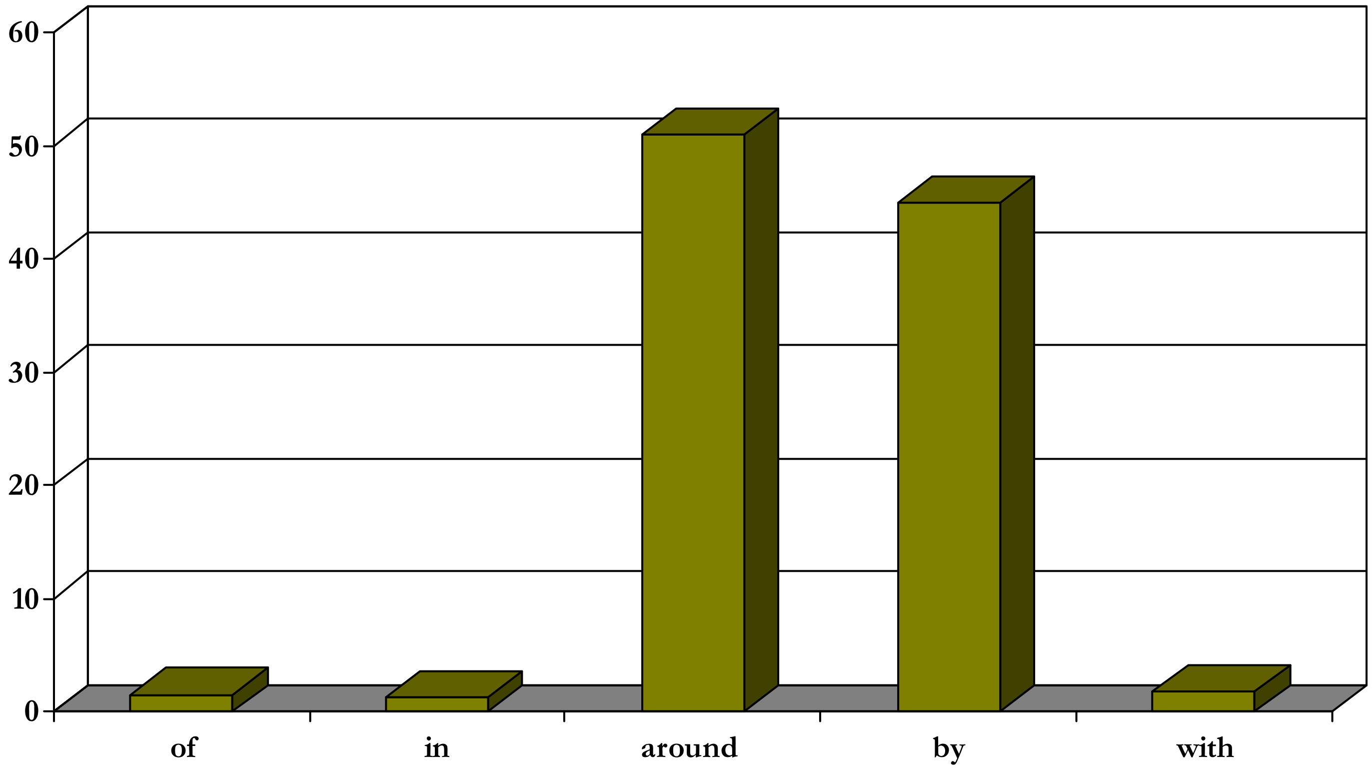
FLAG AS ERROR



“He is fond *with* beer”

Thresholds

FLAG AS OK



"My sister usually gets home *by* 3:00"

Results

Model	Precision	Recall
Lexical	80%	12%
+Combo:tag	82%	14%
+Combo:tag +Extraneous	84%	19%

Typical System Errors

- Noisy context
  - Other errors in vicinity
- Sparse training data
  - Not enough examples of certain constructions
- Biased training data

Related Work

	Method	Performance
[Eeg-Olofsson et al. ’03]	Handcrafted rules for Swedish learners	11/40 prepositions correct
[Izumi et al. ’03, ’04]	ME model to classify 13 error types	25% precision 7% recall
[Lee & Seneff ‘06]	Stochastic model on restricted domain	80% precision 77% recall
[De Felice & Pullman ’08]	ME model (9 prepositions)	~57% precision ~11% recall
[Gamon et al. ’08]	LM + decision trees (12 prepositions)	80% precision

Future Directions

- Noisy Channel Model (MT techniques)
  - Find specific errors or do sentence rewriting
  - [Brockett et al., ‘06; Hermet et al., ‘09]
- Artificial Error Corpora
  - Insert errors into native text to create negative examples
  - [Foster et al., ‘09]
- Test long-range impact of error modules on student writing

Future Directions [WAC '09]

- Current method of training on well-formed text is not error-sensitive:
  - Some errors are more probable than others
    - e.g. "married to" vs. "married with"
  - Different L1's make different types of errors
    - German: "*at* Monday"; Spanish: "*in* Monday"
- These observations are commonly held in the ESL teaching/research communities, but are not captured by current NLP implementations

“Region Web Counts” Approach

- In the absence of a large error-annotated ESL corpus, how does one find common errors?
  - ex: *“married with John” vs. “married to John”
- Novel approach: use region-specific searches to gather data on how different L1’s use certain English constructions
  - Region (or nation) searches = “advanced search”
- Previous work has shown usefulness of web-counts for certain NLP tasks
  - [Lapata & Keller, ’03; Kilgarriff, ‘07]

Web-Counts Example

Region	“depends on”	“depends of”	Ratio
US	92,000,000	267,000	345:1
France	1,500,000	22,700	66:1

*** Counts using Google on March 6, 2009**

- “depends of” is over 5 times more likely to appear in France than in the US
- France’s small ratio may signal a potential error

Summary

- Proof of Concept results appear promising:
  - Showed metric can detect known errors
  - Biasing training data could have a big impact
- Long Range Goal: Automatically determine common errors
  - Run methodology on thousands of constructions
    - Preliminary results on 8500 bigrams appear favorable
  - Add more training data for flagged constructions; determine performance improvement from new model

Conclusions

- Presented a state-of-the-art preposition error detection methodology
  - State-of-the-art preposition selection performance: 79%
  - Accurately detects preposition errors in ESL essays with P=0.84, R=0.19
- This work is included in ETS's *Criterion*SM Online Writing Service and *E-Rater*
- ESL error detection is a growing subfield with a more quickly growing demand
  - Great area for dissertation or project ideas!

Acknowledgments

- Researchers
  - Martin Chodorow [Hunter College of CUNY]
  - Na-Rae Han [University of Pittsburgh]
- Annotators
  - Sarah Ohls [ETS]
  - Waverly Vanwinkle [ETS]
- Other
  - Jill Burstein [ETS]
  - Michael Gamon [Microsoft Research]
  - Claudia Leacock [Butler Hill]

Some More Plugs

- NLP in ETS
  - Postdocs
  - Summer Interns
- 4th Workshop on Innovative Use of NLP for Educational Applications (NAACL-09)
  - http://www.cs.rochester.edu/u/tetreaul/naacl-bea4.html
- NLP/CL Conference Calendar
  - Google "NLP Conferences"
  - http://www.cs.rochester.edu/u/tetreaul/conferences.html