

Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English

Daniel Dahlmeier^{1,2} Hwee Tou Ng^{2,3} Siew Mei Wu⁴

¹SAP Technology and Innovation Platform, SAP Singapore

²NUS Graduate School for Integrative Sciences and Engineering

³Department of Computer Science, National University of Singapore

⁴Centre for English Language Communication, National University of Singapore

BEA8 @ NAACL 2013, Atlanta

Introduction

Motivation

Fact 1:

- Over one billion people in the world are studying English.

Introduction

Motivation

Fact 1:

- Over one billion people in the world are studying English.

Fact 2:

- Computers have made great progress in learning human language thanks to statistical methods.

Introduction

Motivation

Fact 1:

- Over one billion people in the world are studying English.

Fact 2:

- Computers have made great progress in learning human language thanks to statistical methods.

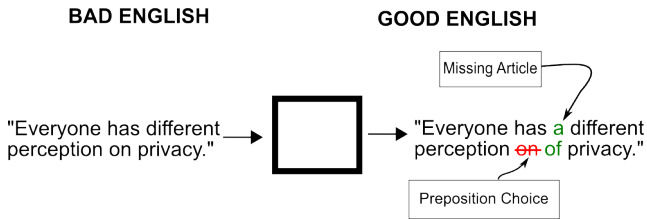
Vision

- Automatic, high-quality grammar correction through statistical NLP.

Statistical approach to grammatical error correction

Basic recipe for statistical grammatical error correction

- 1 Define the confusion set of possible corrections.
- 2 Engineer useful features that are predictive of the correct answer.
- 3 Train a classifier to predict correct word based on context features.
- 4 Test classifier on examples from learner text.



The problem: lack of data

Lack of large, annotated learner corpus

"... 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]

- Statistical approaches require data.
- No large annotated learner corpus for grammatical error correction.
- Existing annotated learner corpora either too small or proprietary.

Work around: train on non-learner text

“Fill in the blank” method

- Create training examples from grammatical non-learner text.
- Take the original word as the class label.
- Extract features from surrounding context.

It's free, no manual annotation required!

Example

Orig: I want to watch **a movie**.

- Class $y = a$
- Features $\mathbf{x} = [\text{left_word}=\text{watch}, \text{right_word}=\text{movie}, \dots]$

Work around: train on artificial learner text

Artificial learner text method

- Create artificial learner errors in non-learner text.
- Take the original word as the class label.
- Extract features from surrounding context and changed word.

Only requires statistics of learner errors.

Example

Orig : I want to watch **a movie**.

Changed : I want to watch **movie**.

- Class $y = a$
- Features $\mathbf{x} = [\text{article}=\text{NULL}, \text{left_word}=\text{watch}, \text{right_word}=\text{movie}, \dots]$

Advantages of real learner data

“Fill in the blank” method — —

- Cannot use the word used by the writer as a feature.
- Errors are not uniformly distributed, removing the word loses information.
- Errors are rare, the original word is often correct!

Artificial learner data —

- Artificial errors might not reflect real errors accurately.
- Generating errors just as hard as correcting them.

Real learner data +

- Learner text is a sample from the real error distribution.
- Allows for analyzing real errors and their distributions.
- Contains multiple, interacting errors.

Outline

In this talk...

- 1 We present the NUS Corpus of Learner English.
- 2 Explain the tag set and annotation process.
- 3 Show statistics of the collected learner data.
- 4 Report on an annotator agreement study for error correction.

NUCLE : NUS Corpus of Learner English

NUS Corpus of Learner English

- About 1,400 essays from university-level students with 1.2 million words.
- Completely annotated with error categories and corrections.
- Annotation performed by English instructors at NUS Centre for English Language Communication (CELC).
- Freely available for research.

Tag set

NUCLE tag set

- 27 error categories grouped into 13 broader categories.
- Developed at NUS Center for English Language Communication.
- Each error annotation contains
 - 1 start and end offset
 - 2 error category
 - 3 correction
 - 4 comment (optional)

Example

- ArtOrDet (Article or Determiner) ... *the technology should not be used in [non-medical — a non-medical] situation.*
- Vform (Verb form) *Will the child blame the parents after he [growing — grows] up?*

Annotation process

Writing, Annotation, and Marking Platform (WAMP)

- **Select** arbitrary, contiguous text spans.
- **Classify** errors by choosing an error tag.
- **Correct** errors by typing the correction into a text box.
- **Comment** to give additional explanations (optional).

Southeast Asia has the oldest and most consistent rainforests on the earth because it is in the equator zone. These forests are very necessary ^{ArtOrDet} for national economies and for the living ^{ArtOrDet} of local population in ^{ArtOrDet} the Southeast Asia. And they are also globally essential requirements in terms of biodiversity and carbon storage. ^{ArtOrDet (Article or Determiner)} of the local early as a result of global demand and expanding economies. These direct causes of deforestation and forest ^{Rloc} degrading are mostly human causes.

Figure: Example of NUCLE corpus.

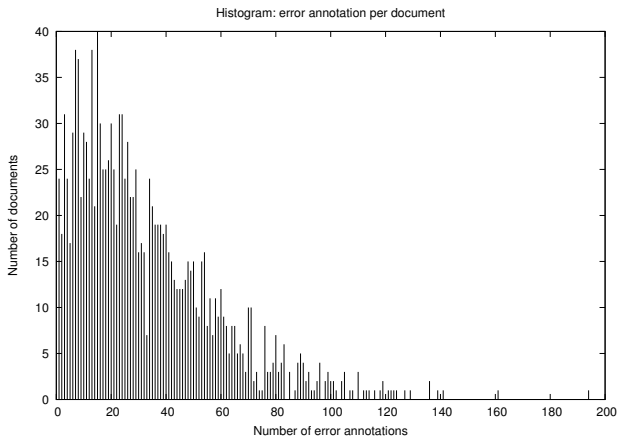
NUCLE corpus statistics

Documents	1,414
Sentences	59,871
Word tokens	1,220,257
Error annotations	46,597
# of error annotations per document	32.95
# of error annotations per 100 word tokens	3.82

NUCLE data collection & annotation

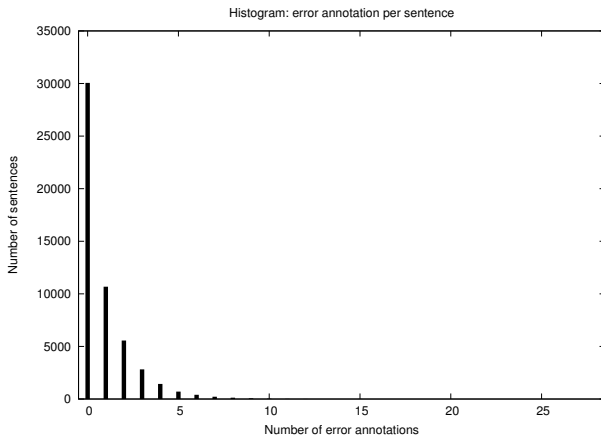
- Essays written by undergraduate students at NUS.
- Essays are take-home assignments for academic writing courses.
- 10 annotators from NUS CELC.

Histogram: errors per document



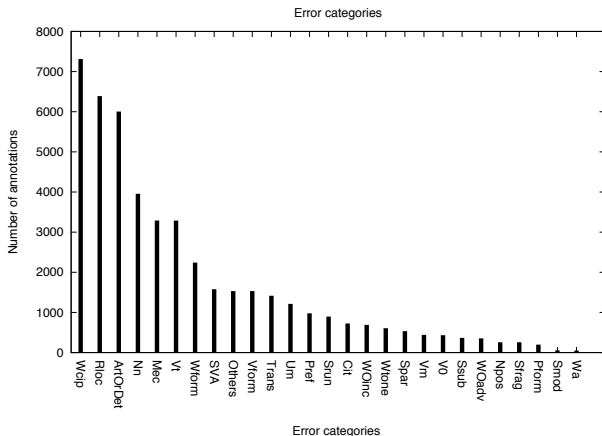
- Skewed distribution.
- Errors are rare in general, but some documents have many errors.

Histogram: errors per sentence



- Most sentences have no or few errors, but some sentences have many errors.

Histogram: error categories



- Some errors are very frequent, many errors are infrequent.
- Article and preposition errors are the most frequent error categories.

Annotator agreement study

Annotator agreement study

- Part of NUCLE pilot study prior to corpus creation.
- 3 annotators from NUS CELC.
- 96 documents.
- Two annotators per document.

Agreement criteria

- **Identification** Is something an error or not?
- **Classification** Agreement of error category, given identification.
- **Exact** Agreement of error category and correction, given identification.

Identification agreement

Source	:	This phenomenon opposes the real .
Annotator A	:	This phenomenon opposes the real .
Annotator B	:	This phenomenon opposes the real .

- Agree that *real* is an error.
- Disagree whether *the* is an error.
- Agree that all other tokens are correct.

Kappa agreement between two annotators

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

$$Pr(a) = \frac{\# \text{agreed tokens}}{\# \text{total tokens}}$$

$$Pr(e) = Pr(A = 1)Pr(B = 1) + Pr(A = 0)Pr(B = 0)$$

$$Pr(A = 1) = \frac{\# \text{ annotated as error by annotator A}}{\# \text{ total tokens}}$$

Classification and exact agreement

Source : This phenomenon opposes the real .

Annotator A : This phenomenon opposes the (real → reality (Wform)) .

Annotator B : This phenomenon opposes the (real → reality (Wform)) .

- Only consider tokens where annotators agree that they are errors.
- **Classification:** Agree that *real* is word form error.
- **Exact:** Agree that *real* is word form error and should be corrected as *reality*.

Results

Annotators	Kappa-iden	Kappa-class	Kappa-exact
A – B	0.4775	0.6206	0.5313
A – C	0.3627	0.5352	0.4956
B – C	0.3230	0.4894	0.4246
Average	0.3877	0.5484	0.4838

Results

- Fair agreement for identification.
- Moderate agreement for classification and exact agreement.
- Identifying errors seems to be harder than classifying or correcting them.
- Error correction is difficult, especially detecting errors.

Related work

Learner corpora for NLP

- ICLE - International Corpus of Learner English.
- Chinese Learner English Corpus.
 - Rozovskaya and Roth produced annotations for about 63,000 words from both corpora [Rozovskaya and Roth2010].
- CLC - Cambridge Learner Corpus [Yannakoudakis et al.2011].
- HOO 2011 and HOO 2012 shared task [Dale et al.2012].
- CoNLL 2013 shared task [Ng et al.2013].

Conclusion

Conclusion

- The NUS Corpus of Learner English is a one-million word learner corpus.
- Contains annotations of error categories and corrections.
- Error correction is a difficult problem, even for humans.

References

References I



D. Dahlmeier and H.T. Ng.
2012.

Better evaluation for grammatical error correction.
In *Proceedings of HLT-NAACL*, pages 568–572.



R. Dale, I. Anisimoff, and G. Narroway.
2012.




HOO 2012: A report on the preposition and determiner error correction shared task.
In *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–62.



C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault.
2010.

Automated Grammatical Error Detection for Language Learners.
Morgan & Claypool Publishers.

References II

-  H.T. Ng, S.M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault.
2013.
The CoNLL-2013 shared task on grammatical error correction.
In *To appear in Proceedings of the Seventeenth Conference on Computational Natural Language Learning*.
-  A. Rozovskaya and D. Roth.
2010.
Training paradigms for correcting errors in grammar and usage.
In *Proceedings of HLT-NAACL*, pages 154–162.
-  H. Yannakoudakis, T. Briscoe, and B. Medlock.
2011.
A new dataset and method for automatically grading ESOL texts.
In *Proceedings of ACL:HLT*, pages 180–189.