Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics

Paper Review

Yue Yu

March 15, 2016

for Topics in Natural Language Processing (INFR11113)

Table of contents

Backgrounds

Function Word Vs. Content Word

Errors in function words (e.g. articles or prepositions)

I am 0/a student.*

We would consider {a, an, the} as possible corrections for the missing article.

Last October, I came in/to Tokyo.*

To correct this preposition, we would consider the most frequent prepositions {on, from, for, of, about, to, at, with, by}

Function Word Vs. Content Word

Errors in content words

- Now I felt a big anger. → great anger [confused via meaning]
- It includes articles over ancient Greek sightseeings as the Alcropolis or other famous places. → ancient sites [confused via form]
- Deep regards, John Smith → kind regards [(seemingly) unrelated]
- The company had great turnover, which was noticable in this market. → high turnover [context-dependent interpretation]

The Task

Capturing Anomalies in the Choice of Content Words

Previous Approach

1. Search for the most suitable correction among the alternatives typically composed of **synonyms**, **homophones** or **L1-related paraphrases**.
2. This approach compares original word combinations to their alternatives using corpus statistics, where low frequency or low collocational strength clearly signifies an error.
3. Detection and correction can occur simultaneously.

New Challenges

- Language learners are creative in their writing (many of the combinations are corpus-unattested);
- Learners might be misled and confused if they are frequently notified by a system that something is an error when it is not (falsely identified errors are more harmful for language learning than missed errors).

Errors in content words

- Now I felt a big anger. → great anger [confused via meaning]
- It includes articles over ancient Greek sightseeings as the Alcropolis or other famous places. → ancient sites [confused via form]
- Deep regards, John Smith → kind regards [(seemingly) unrelated]
- The company had great turnover, which was noticable in this market. → high turnover [context-dependent interpretation]

Solution

Compositional Distributional Semantics

Methodology

Data Annotation

The data for training and testing is annotated in 3 steps:

1. A list of 61 adjectives that are most problematic (typical errors) for learners is compiled from CLC-FCE dataset [5].
2. Using this set of 61 adjectives, we extracted AN combinations from the Cambridge Learner Corpus (CLC).
3. Based on the British National Corpus (BNC), we select the corpus-unattested (previously unseen in corpus) AN combinations. We have compiled a set of 798 AN combinations.

Data Annotation

We also distinguish between out-of-context (OOC) and in-context (IC) annotation.

OOC Annotation

considered out of their original context of use

IC Annotation

only considered in their original context of use

Distributional Semantic Models

- Key assumption: word meaning can be approximated by a words distribution
- Method: represent words with distributional vectors, dimensions = co-occurrence with context words
- Hypothesis: semantically similar words occur in similar contexts

Compositional Models

- additive (add) [3]
- multiplicative (mult) [3]
- adjective-specific linear maps (alm) [1]

The first two models are symmetric. While, in the alm model, adjectives are functions (weight matrices) mapping from noun meanings to a composite noun-like vector for the ANs

Alm Model

OLD	*bloom*	*buy*
bloom	10	0
buy	6	15

$\times$

	tree
bloom	34
buy	10

$=$

	OLD(tree)
bloom	$(10 \times 34) + (0 \times 10) = 340$
buy	$(6 \times 34) + (15 \times 10) = 354$

Feature Selection

Several semantic measures (1 to 8) for detecting semantic anomaly have been introduced in previous work [4][2].

1. Vector length (VLen)
2. Cosine to the input noun (cosN)
3. Cosine to the input adjective (cosA)
4. Density of the neighbourhood populated by 10 nearest neighbours (dens)
5. Density among the 10 nearest neighbours (densAll)
6. Ranked density in close proximity (Rdens)
7. Number of neighbours within close proximity (num)
8. Overlap between the 10 nearest neighbours and constituent noun/adjective (OverAN)

Feature Selection

Some additional measures (9 to 13) are also added to help distinguish between correct and incorrect word combinations:

9. Overlap between the 10 nearest neighbours and input noun (OverN)
10. Overlap between the 10 nearest neighbours and input adjective (OverA)
11. Overlap between the 10 nearest neighbours for the AN and constituent noun/adjective (NOverAN)
12. Overlap between the 10 nearest neighbours for the AN and input noun (NOverN)
13. Overlap between the 10 nearest neighbours for the AN and input adjective (NOverA)

Classification

Baseline System

A system similar to the previous approach.

Supervised Classifier

- The best results so far have been obtained with the Decision Tree classifier using NLTK,
- with 5-fold cross-validation on 798 ANs.

Result

Evaluation on Individual Features

Metric	*add*	*mult*	*alm*
VLen	0.7589	0.7690	0.1676
cosN	0.1621	**0.0248**	**0.0227**
cosA	**0.0029**	0.4782	0.0921
dens	0.6731	0.1182	0.1024
densAll	0.4967	0.1026	0.1176
RDens	0.2786	0.8754	0.1970
num	0.3132	0.4673	0.3765
OverAN	0.8529	0.1622	0.2808
OverA	**0.0151**	0.6377	0.4886
OverN	**0.0138**	0.0764	0.4118
NOverAN	0.3941	0.6730	0.0858
NOverA	**0.0009**	0.3342	0.1575
NOverN	**0.0018**	0.1463	0.1497

Table 2: *p* values, *out-of-context* annotation

Metric	*add*	*mult*	*alm*
VLen	0.6675	**0.0027**	**0.0111**
cosN	**0.0417**	**0.0070**	0.1845
cosA	**0.00003**	0.1791	0.1442
dens	0.4756	0.7120	0.1278
densAll	0.2262	0.7139	0.5310
RDens	0.8934	0.8664	0.1985
num	0.7077	0.7415	0.4259
OverAN	0.1962	0.8635	0.5669
OverA	**0.00007**	0.7271	0.6229
OverN	**0.0017**	0.9680	0.7733
NOverAN	**0.0227**	0.3473	0.1587
NOverA	**0.000004**	0.3749	0.1576
NOverN	**0.0001**	0.6651	0.2610

Table 3: *p* values, *in-context* annotation

p value represents statistical significance of the difference between the groups of correct and incorrect ANs: the lower the better.

Performance of Classifier

Type	Accuracy	Baseline	LB	UB
OOC	**0.8113** ± 0.0149	0.3897	0.7932	0.8650
IC	**0.6535** ± 0.0189	0.5147	0.5063	0.7467

Table 4: *Decision Tree* classification results

Type	*P* (correct)	*P* (incorrect)
OOC	0.8193	0.7500
IC	0.6241	0.6850

Table 5: Classification precision

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$$

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

Conclusion

Summary

- This paper presented an annotated dataset of learner errors in AN, which contains examples not seen in a native corpus of English (Challenge).
- Error detection is casted as a binary classification task and a supervised classifier that uses semantically-motivated features is implemented (Solution).
- The best results are obtained with a Decision Tree classifier and the resulting error detection system can identify errors with high precision and accuracy (Result).

Questions?

References I

M. Baroni and R. Zamparelli.
Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space.
In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics, 2010.

E. Kochmar and T. Briscoe.
Capturing anomalies in the choice of content words in compositional distributional semantic space.
In *RANLP*, pages 365–372, 2013.

J. Mitchell and M. Lapata.
Vector-based models of semantic composition.
In *ACL*, pages 236–244, 2008.

References II

E. M. Vecchi, M. Baroni, and R. Zamparelli.
(linear) maps of the impossible: capturing semantic anomalies in distributional space.
In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9. Association for Computational Linguistics, 2011.

H. Yannakoudakis, T. Briscoe, and B. Medlock.
A new dataset and method for automatically grading esol texts.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics, 2011.