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

Ekaterina Kochmar and Ted Briscoe

Computer Laboratory, University of Cambridge, UK

COLING 2014

Outline

Motivation

- Growing interest in error detection and correction (EDC)
  - Growing number of non-native speakers of English
  - Growing number of conference papers, books and tutorials on this task
  - Shared tasks on grammatical EDC (Dale and Kilgarriff, 2011; Dale *et al.*, 2012; Ng *et al.*, 2013, 2014)
- Most often focus on **function words**
  - Most frequent error types – should be addressed by any EDC system
  - Closed class words with finite sets of confusions
  - Recurrent errors
- Less on **content words**
  - Third most frequent error type (Leacock *et al.*, 2010)
  - Open class words with unlimited sets of confusions
  - Convey meaning

Errors in Function Words

Example

I am $\emptyset$*/*a* student.

- Possible corrections: $\{$*a, an, the*$\}$
- Recurrent: *I am* + occupation
- Contexts: highly informative, can be used to extract features
- Treated as a 4-class classification problem: $\{\emptyset$, *a, an, the*$\}$
- Machine learning-based approaches

Errors in Content Word Combinations

Examples of errors in adjective–noun combinations

- Similar in **meaning**: Now I felt a big anger. → great anger
- Similar in **form**: It includes articles over ancient Greek sightseeings as the Alcropolis or other famous places. → ancient sites
- **Not obvious**: Deep regards, John Smith → kind regards
- **Context-dependent** interpretation: The company had great turnover, which was noticable in this market. → high turnover

Errors in content words vs errors in function words

- Possible corrections: depend on the original combination
- Reasons for confusion: more diverse
- Contexts: more diverse, less informative
- Classification approach: how many classes?
- Often result in **semantically anomalous** word combinations

Contributions of this Work

Focus

Error detection in adjective–noun (AN) combinations

Contributions

- present and release an error-annotated AN dataset extracted from learner data
- show how compositional distributional semantic models can be applied to detect semantic anomalies in this dataset
- demonstrate that the output of these models can be used to derive features for error detection in AN combinations

Contributions of this Work

Focus

Error detection in adjective–noun (AN) combinations

Contributions

- present and release an error-annotated AN dataset extracted from learner data
- show how compositional distributional semantic models can be applied to detect semantic anomalies in this dataset
- demonstrate that the output of these models can be used to derive features for error detection in AN combinations

Contributions of this Work

Focus

Error detection in adjective–noun (AN) combinations

Contributions

- present and release an error-annotated AN dataset extracted from learner data
- show how compositional distributional semantic models can be applied to detect semantic anomalies in this dataset
- demonstrate that the output of these models can be used to derive features for error detection in AN combinations

AN Dataset: Objectives

Collect AN combinations that

1. exemplify **typical** errors committed by language learners in the choice of content words
2. are **challenging** for an EDC system

Data Collection

To exemplify **typical** errors

- examined the publicly available CLC-FCE dataset (Yannakoudakis *et al.*, 2011)
- analysed errors in AN combinations committed by language learners using the error annotation (Nicholls, 2003)
- compiled a list of 61 adjectives that are **most problematic** for learners

To collect examples **challenging** for an EDC system

- extracted AN combinations from the Cambridge Learner Corpus (CLC)
- focused on AN combinations previously **unseen** in a native English corpus (BNC)

Data Collection

Why unattested combinations are challenging for an EDC algorithm?

- cannot be effectively handled with simple comparison-based approaches
- language learners are creative $\Rightarrow$ there is a substantial number of previously unseen combinations
- no corpus could cover all possible acceptable content word combinations in language

Annotation Scheme

798 AN combinations extracted from the CLC

Distinguish between **out-of-context (OOC)** and **in-context (IC)** annotation

classic dance?

- **OOC** – correct: *They performed a classic Ceilidh dance.*
- **IC** – most often incorrect: *I have tried a rock'n'roll dance and a classic*|classical dance already.*

Annotate AN combinations for error *location* (adj/noun/both) and *source*:

- Semantically related words: big*|long history, large*|broad knowledge
- Form-related words: classic*|classical dance, economical*|economic crisis
- Other (not related) confusion: clear*|clever people, deep*|great majesty

Annotation Examples

C-J-N

Correct both out-of-context and in-context

Example: I found a *great cinema* for us tonight.

C-JF-N

Correct out-of-context
Incorrect in-context due to a form-related confusion

Example: I have tried a rock'n'roll dance and a *classic*|*classical dance* already.

I-JS-NN

Incorrect both out-of-context and in-context.
Semantically related confusion on the adjective + confusion on the noun

Example: This *strong*|*strict education*|*upbringing* made me very self-confident and proud.

Annotation Examples

C-J-N

Correct both out-of-context and in-context

Example: I found a *great cinema* for us tonight.

C-JF-N

Correct out-of-context
Incorrect in-context due to a form-related confusion

Example: I have tried a rock'n'roll dance and a *classic*|*classical dance* already.

I-JS-NN

Incorrect both out-of-context and in-context.
Semantically related confusion on the adjective + confusion on the noun

Example: This *strong*|*strict education*|*upbringing* made me very self-confident and proud.

Annotation Examples

C-J-N

Correct both out-of-context and in-context

Example: I found a *great cinema* for us tonight.

C-JF-N

Correct out-of-context
Incorrect in-context due to a form-related confusion

Example: I have tried a rock'n'roll dance and a *classic*|*classical dance* already.

I-JS-NN

Incorrect both out-of-context and in-context.
Semantically related confusion on the adjective + confusion on the noun

Example: This *strong*|*strict education*|*upbringing* made me very self-confident and proud.

Data Annotation

100 examples extracted randomly and annotated by 4 annotators

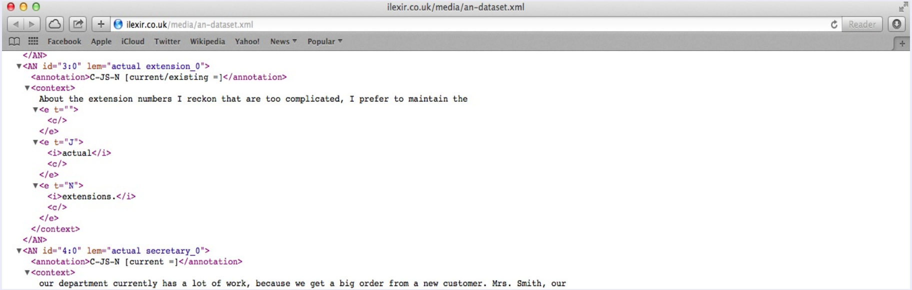
Annotation	OOC	IC
Agreement	**0.8650** $\pm$ 0.0340	**0.7467** $\pm$ 0.0221
Cohen's *kappa*	0.6500 $\pm$ 0.0930 (*substantial*)	0.4917 $\pm$ 0.0463 (*moderate*)

Table : Average observed agreement and kappa values.

OOC	IC
79.32% correct (`C`) 20.68% incorrect (`I`)	**50.63**% correct (`C-J-N`) 49.37% incorrect (other)

Table : Distribution of correct and incorrect instances.

Dataset Release



http://ilexir.co.uk/applications/adjective-noun-dataset/

Previous Approaches to EDC in Content Words

Previous approaches

- **Error correction** for already detected errors (Liu *et al.*, 2009; Dahlmeier and Ng, 2011)
- **Writing improvement** (Chang *et al.*, 2008; Futagi *et al.*, 2008):
  - for each combination X, check for more fluent/native-like alternatives Y
  - compare alternatives Y to X using some frequency-based measure
  - if $\exists\ Y_i$ more fluent than $X \Rightarrow X$ is an error, Y_i is a correction

Baseline system implementation

- collect the sets of alternatives for adjectives and nouns using WordNet
  - adjectives=$\{$*original, synonyms*$\}$
  - nouns=$\{$*original, synonyms*$\}$ or $\{$*original, synonyms, hyper-/hyponyms*$\}$
- cross the sets of alternatives: adjectives $\cap$ nouns
- select the alternative with the highest collocational strength
- if selected alternative $\neq$ original, detect an error

Baseline System

Collocational strength

Normalized pointwise mutual information (*npmi*) of an *an* combination

$$npmi(a, n) = \frac{pmi(a, n)}{-log[p(a, n)]} \quad (1) \qquad pmi(a, n) = log\,\frac{p(a, n)}{p(a)p(n)} \quad (2)$$

Accuracy

Proportion of correctly identified correct (TN) and incorrect (TP) AN combinations

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

Upper (UB) and lower (LB) bounds

UB = observed inter-annotator agreement

LB = majority class baseline

Baseline System: Results

Results

Type	Baseline	LB	UB
OOC	**0.3897**	0.7932	0.8650
IC	**0.5147**	0.5063	0.7467

Table : Baseline System

Limitations

- System aimed at finding the most fluent alternative
  $\Rightarrow$ *any* corpus-attested alternative better than the corpus-unattested original
- Overcorrection (*false positives*):
  important conversation corrected to *serious conversation*
- Lack of semantically motivated decisions (*false negatives*):
  **high shyness* not detected as no alternative found

Compositional Distributional Semantic Models for EDC

Advantages

- Many errors stem from **semantic** mismatch:
  incorrect content word combinations $\sim$ anomalous combinations
- Compositional distributional semantic models do not rely directly on corpus statistics $\Rightarrow$ can be applied to previously unseen combinations
- Promising results on related tasks:
  - semantic anomaly detection (Vecchi *et al.*, 2011)
  - tests on learner data (Kochmar and Briscoe, 2013)

Objective

Show how the output of the compositional distributional semantic models can be used as features in a classifier

Semantic Space Construction

Source corpus

- British National Corpus
- Lemmatised, tagged and parsed with the RASP system (Briscoe *et al.*, 2006)
- Statistics extracted at the lemma level, no inflectional information

Semantic space

- Target words and combinations:
  - $\sim$ 8K nouns (most frequent in the corpus + test ones)
  - $\sim$ 4K adjectives (most frequent in the corpus + test ones)
  - $\sim$ 64K ANs with >100 occurrences in the corpus
- Context words:
  - 10K most frequent nouns, adjectives and verbs
  - Co-occurrence counts converted into Local Mutual Information scores (Evert, 2005)
- The original $76K \times 10K$ matrix reduced to $76K \times 300$ using SVD

Models of Semantic Composition

Additive and multiplicative models (Mitchell and Lapata, 2008)

Component-wise vector addition and multiplication:

$$c_i = a_i + b_i \qquad\qquad c_i = a_i \times b_i$$

Adjective–specific linear maps (Baroni and Zamparelli, 2010)

- Nouns represented by their distributional vectors
- Adjectives are matrices encoding distributional functions:
  new in *new friend* $\neq$ *new* in *new shoes*
  $\Rightarrow$ *new friend* $= \mathcal{NEW}(friend)$, *new shoes* $= \mathcal{NEW}(shoes)$
- Matrices learned from data using regression
- AN vector derived by matrix-by-vector multiplication:
  $\mathcal{ADJ}(noun) = \mathbf{F}_{adj} \times \overrightarrow{noun} = \overrightarrow{AN}$

Measures of Semantic Anomaly

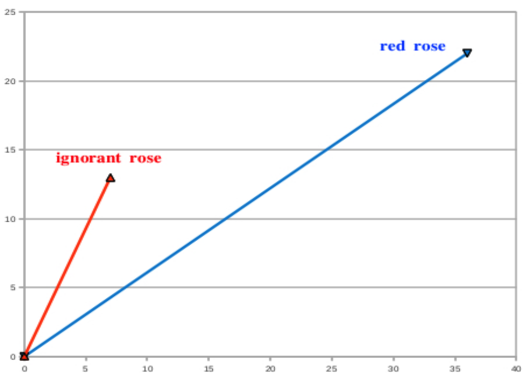
13 measures of semantic anomaly

- Length-based (1):
  - *Vector length*
- Distance to component words (2):
  - *Cosine to the input noun*
  - *Cosine to the input adjective*
- Neighbourhood-based (10):
  - *Density of the neighbourhood populated by* 10 *nearest neighbours*
  - *Overlap between the* 10 *nearest neighbours and constituent noun/adjective*
  - *Overlap between the* 10 *nearest neighbours and neighbours of the constituent noun/adjective*

Measures of Semantic Anomaly: Vector Length

Example: Vector length

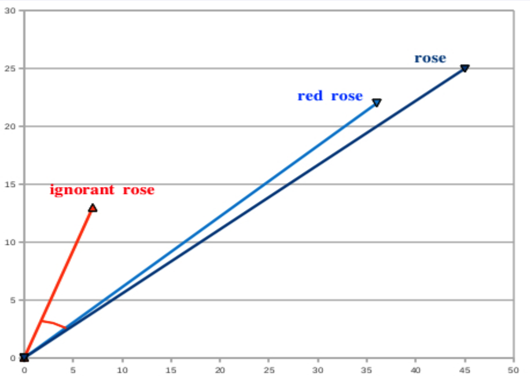
In anomalous/incorrect ANs, the counts in the input vectors are distributed differently
→ some "incompatible dimensions" would receive low counts
→ anomalous AN vectors are expected to be shorter



Measures of Semantic Anomaly: Distance to Components

Example: Cosine to the component noun

Anomalous/incorrect ANs are less similar to the input nouns
→ their vectors are expected to have lower cosine to the input noun vector



Measures of Semantic Anomaly: Neighbourhood-based

- **Example 1**: *Neighbourhood density*:
  Semantically acceptable/correct ANs are expected to have denser neighbourhoods, and anomalous/incorrect AN vectors – to have sparser neighbourhoods
  (measured as an average cosine/distance to the 10 nearest neighbours)
- **Example 2**: *Component overlap*:
  Semantically acceptable/correct ANs are expected to be placed in the neighbourhoods populated by similar words and combinations
  (measured as a proportion of neighbours among 10 nearest ones containing the same constituent words as in the tested AN)

red rose	*ignorant rose*
(x) **rose**	people
red (x)	blind people
flower	like-minded
...	...

Evaluation

Approach

- For the measures of semantic anomaly, compute the difference between the mean values for the vectors for correct and incorrect ANs (Vecchi *et al.*, 2011, Kochmar and Briscoe, 2013)
- Apply *t*-test, statistical significance level $p < 0.05$
- Test an ability of the measures to distinguish the correct ANs from the incorrect ones in general

Results

- Showed that most of the measures distinguish between correct and incorrect examples with at least one of the models
- Confirmed that they can be used as features

Machine Learning Approach

General framework

- Treat error detection in content words as a binary classification problem
- Apply an ML classifier
- Use the values of the semantic measures as features

Implementation

- Applied 5-fold cross-validation, with 80% training and 20% testing
- *Decision Tree* classifier using *NLTK* (Bird et al., 2009)
- Feature binning used: 10 value intervals for each feature
- 14 feature types:
  - values in the range $[-1, 1]$ (i.e., *VLen* normalised)
  - adjective identity used as a feature: e.g., ANs with an adjective adj_1 might have higher *cosN* values than ANs with an adjective adj_2

Semantical System: Results

Results

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

Table : *Decision Tree* classification results

Missed errors

Most cases – semantically related confusion:

e.g., *big**|*great anger*, *biggest**|*greatest painter*, *small**|*short speech*

Analysis and Discussion

Precision of the EDC algorithms

- High precision to facilitate language learning (Nagata and Nakatani, 2010)
- Falsely identified errors mislead learners

$$P = \frac{TP}{TP + FP} \tag{4}$$

$\Rightarrow$ if $P < 0.5$ on errors, the system tags correct instances as errors more frequently than it correctly detects errors

Precision

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

Table : Classification precision

Conclusions

Summary

- Presented and released an error-annotated AN dataset extracted from learner data
- Showed how compositional distributional semantic models can be applied to detect semantic anomalies in this dataset
- Implemented a classifier that uses semantically motivated features and shows good precision and accuracy

Future work

- Extend the system to perform error correction
- Implement an EDC system for other types of content word combinations

Thank you!

Dataset available at:
http://ilexir.co.uk/applications/adjective-noun-dataset/

Contact: Ekaterina.Kochmar@cl.cam.ac.uk

References

M. Baroni and R. Zamparelli, 2010. *Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space*. In Proceedings of the EMNLP-2010

S. Bird, E. Klein, and E. Loper, 2009. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. O'Reilly Media

T. Briscoe, J. Carroll and R. Watson, 2006. *The Second Release of the RASP System*. In Proceedings of the COLING/ACL-2006 Interactive Presentation Sessions

Y.-C. Chang, J. S. Chang, H.-J. Chen and H.-C. Liou, 2008. *An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology*. Computer Assisted Language Learning, 21(3)

D. Dahlmeier and H. T. Ng, 2011. *Correcting Semantic Collocation Errors with L1-induced Paraphrases*. In Proceedings of the EMNLP-2011

R. Dale and A. Kilgarriff, 2001. *Helping Our Own: The HOO 2011 Pilot Shared Task*. In Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), volume Helping Our Own: The HOO 2011 Pilot Shared Task

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 7th Workshop on Innovative Use of NLP for Building Educational Applications

S. Evert, 2005. *The Statistics of Word Cooccurrences*. PhD thesis, Stuttgart University

Y. Futagi, P. Deane, M. Chodorow, and J. Tetreault, 2008. *A computational approach to detecting collocation errors in the writing of non-native speakers of English*. Computer Assisted Language Learning, 21(4)

E. Kochmar and T. Briscoe, 2013. *Capturing Anomalies in the Choice of Content Words in Compositional Distributional Semantic Space*. In Proceedings of the Recent Advances in Natural Language Processing (RANLP-2013)

C. Leacock, M. Chodorow, and J. Tetreault, 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool Publishers

A. L.-E. Liu, D. Wible, and N.-L. Tsao, 2009. *Automated suggestions for miscollocations*. In Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications

J. Mitchell and M. Lapata, 2008. *Vector-based models of semantic composition*. In Proceedings of ACL

H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, 2013. *The CoNLL-2013 Shared Task on Grammatical Error Correction*. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task

H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant, 2014. *The CoNLL-2014 Shared Task on Grammatical Error Correction*. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task

D. Nicholls, 2003. *The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT*. In Proceedings of the Corpus Linguistics conference

E. Vecchi, M. Baroni, and R. Zamparelli, 2011. *(Linear) maps of the impossible: Capturing semantic anomalies in distributional space*. In Proceedings of the DISCO (Distributional Semantics and Compositionality) Workshop at ACL 2011

H. Yannakoudakis, T. Briscoe, and B. Medlock, 2011. *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