Automated Suggestions for Miscollocations

Anne Li-E Liu
Research Centre for English and Applied Linguistics
University of Cambridge
Cambridge, CB3 9DP,
United Kingdom
lel29@cam.ac.uk

David Wible
Graduate Institute of Learning and Instruction
National Central University
Jhongli City, Taoyuan County
32001, Taiwan
wible45@yahoo.com

Nai-Lung Tsao
Graduate Institute of Learning and Instruction
National Central University
Jhongli City, Taoyuan County
32001, Taiwan
beaktsao@gmail.com

Abstract

One of the most common and persistent error types in second language writing is collocation errors, such as *learn knowledge* instead of *gain* or *acquire knowledge*, or *make damage* rather than *cause damage*. In this work-in-progress report, we propose a probabilistic model for suggesting corrections to lexical collocation errors. The probabilistic model incorporates three features: word association strength (MI), semantic similarity (via WordNet) and the notion of shared collocations (or intercollocability). The results suggest that the combination of all three features outperforms any single feature or any combination of two features.

1 Collocation in Language Learning

The importance and difficulty of collocations for second language users has been widely acknowledged and various sources of the difficulty put forth (Granger 1998, Nesselhauf 2004, Howarth 1998, Liu 2002, inter alia). Liu's study of a 4-million-word learner corpus reveals that verb-noun (VN) miscollocations make up the bulk of the lexical collocation errors in learners' essays. Our study focuses, therefore, on VN miscollocation correction.

2 Error Detection and Correction in NLP

Error detection and correction have been two major issues in NLP research in the past decade. Projects involving learner corpora in analyzing and categorizing learner errors include NICT Japanese Learners of English (JLE), the Chinese Learners of English Corpus (Gamon et al., 2008) and English Taiwan Learner Corpus (or TLC) (Wible et al., 2003). Studies that focus on providing automatic correction, however, mainly deal with errors that derive from closed-class words, such as articles (Han et al., 2004) and prepositions (Chodorow et al., 2007). One goal of this work-in-progress is to address the less studied issue of open class lexical errors, specifically lexical collocation errors.

3 The Present Study

We focus on providing correct collocation suggestions for lexical miscollocations. Three features are employed to identify the correct collocation substitute for a miscollocation: word association measurement, semantic similarity between the correction candidate and the misused word to be replaced, and intercollocability (i.e., the concept of shared collocates in collocation clusters proposed by Cowie and Howarth, 1995). NLP research on learner errors includes work on error detection and error correction. While we are working on both, here we report specifically on our work on lexical miscollocation correction.

4 Method

We incorporate both linguistic and computational perspectives in our approach. 84 VN miscollocations from Liu's (2002) study were employed as the training and the testing data in that each comprised 42 randomly chosen miscollocations. Two experienced English teachers[1] manually went through the 84 miscollocations and provided a list of correction suggestions. Only when the system output matches to any of the suggestions offered

[1] One native speaker and one experienced non-native English teacher.

by the two annotators would the data be included in the result. The two main knowledge resources that we incorporated are British National Corpus[2] and WordNet (Miller, 1990). BNC was utilized to measure word association strength and to extract shared collocates while WordNet was used in determining semantic similarity. Our probabilistic model that combines the features is described in sub-section 4.4. Note that all the 84 VN miscollocations are combination of incorrect verbs and focal nouns, our approach is therefore aimed to find the correct verb replacements.

4.1 Word Association Measurement

The role of word association in miscollocation suggestions are twofold: 1. all suggested correct collocations in any case have to be identified as collocations; thus, we assume candidate replacements for the miscollocate verbs must exceed a threshold word association strength with the focal noun; 2. we examine the possibility that the higher the word association score the more likely it is to be a correct substitute for the wrong collocate. We adopt Mutual Information (Church et al. 1991) as our association measurement.

4.2 Semantic Similarity

Both Gitsaki et al. (2000) and Liu (2002) suggest a semantic relation holds between a miscollocate and its correct counterpart. Following this, we assume that in the 84 miscollocations, the miscollocates should stand in more or less a semantic relation with the corrections. For example, *say* in an attested learner miscollocation *say story* is found to be a synonym of the correct verb *tell* in WordNet. Based on this assumption, words that show some degree of semantic similarity with the miscollocate are considered possible candidates for replacing it. To measure similarity we take the synsets of WordNet to be nodes in a graph. We quantify the semantic similarity of the incorrect verb in a miscollocation with other possible substitute verbs by measuring graph-theoretic distance between the synset containing the miscollocate verb and the synset containing candidate substitutes. In cases of polysemy, we take the closest synsets for the distance measure. If the miscollocate and the candidate substitute occur in the same synset, then the distance between them is zero.

The similarity measurement function is as follows (Tsao et al., 2003):

$$sim(w_1, w_2) = \max_{s_i \in synset(w_1), s_j \in synset(w_2)} (1 - \frac{dis(s_i, s_j)}{2 \times \max(L_{s_i}, L_{s_j})})$$

,where $dis(s_i, s_j)$ means the node path length between the synset s_i and s_j in WordNet hyper/hypo tree. L_s means the level number of s in hyper/hypo tree and the level of top node is 1. Multiplying $\max(L_{s_i}, L_{s_j})$ by 2 ensures the similarity is less than 1. If s_i and s_j are synonymous, the similarity will be 1.

4.3 Shared Collocates in Collocation Clusters

Futagi et al (2008) review several studies which adopt computational approaches in tackling collocation errors; yet none of them, including Futagi et al., include the notion of collocation cluster. We borrow the cluster idea from Cowie & Howarth (1995) who propose 'overlapping cluster' to denote sets of collocations that carry similar meaning and shared collocates. Figure 1 represents a collocation cluster that expresses the concept of 'bringing something into actuality.' The key here is that not all VN combinations in Figure 1 are acceptable. While *fulfill* and *achieve* collocate with the four nouns on the right, *realize* does not collocate with *purpose*, as indicated by the dotted line. Cowie and Howarth's point is that collocations that can be clustered via overlapping collocates can be the source of collocation errors for language learners. That both *fulfill* and *reach* collocate with *goal* and the further collocability of *fulfill* with *ambition* and *purpose* plausibly lead learners to assume that *reach* shares this collocability as well, leading by overgeneralization to the miscollocations *reach an ambition* or *reach a purpose*.

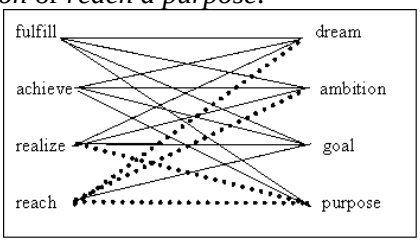

Figure 1. Collocation cluster of 'bringing something into actuality'

[2] *The British National Corpus*, version 3 (BNC XML Edition). 2007. URL: http://www.natcorp.ox.ac.uk/

We employ the ideas of 'collocation cluster' and 'shared collocates' in identifying correct counterparts to the miscollocations. Specifically, taking the miscollocation *reach their purpose* as a starting point, our system generates a collocation cluster by finding the verbs that collocate with *purpose* and nouns that *reach* collocates with. We consider this formed cluster the source that contains the possible correct replacement for *reach* in *reach their purpose*. By finding verbs that not only collocate with *purpose* but also share the most other collocating nouns with the wrong verb *reach*, successfully, we identified candidate substitutes *fulfill* and *achieve* for the incorrect verb *reach.*

4.4 Our Probabilistic Model

The three features we described above are integrated into a probabilistic model. Each feature is used to look up the correct collocation suggestion for a miscollocation. For instance, *cause damage*, one of the possible suggestions for the miscollocation *make damage*, is found to be ranked the 5th correction candidate by using word association measurement merely, the 2nd by semantic similarity and the 14th by using shared collocates. If we combine the three features, however, *cause damage* is ranked first.

The conditional probability of the case where the candidate is a correct one can be presented as:

$$P\ (c \text{ is a correct verb } \left| F_{c,m} \right.)$$

where c means a candidate for a specific miscollocation and $F_{c,m}$ means the features values between m (misused words) and c (candidates). According to Bayes theorem and Bayes assumption, which assume that these features are independent, the probability can be computed by:

$$P(S_c \left| F_{c,m} \right.) = \frac{P(F_{c,m} \left| S_c \right.) P(S_c)}{P(F_{c,m})} \approx \frac{\prod_{f \in F_{c,m}} P(f \left| S_c \right.) P(S_c)}{\prod_{f \in F_{c,m}} P(f)}$$

where S_c means the situation 'c is a correct verb', as described above and f is one of the three particular features. We use probability values to choose and rank the K-best suggestions.

5 Experimental Results

Any found VN combination via our probabilistic approach was compared to the suggestions made by the two human experts. A match would be counted as a true positive. A discrete probability distribution is produced for each feature. We divided feature value into five levels and obtained prior predicting value for each level of the three features. For example, we divided MI value to five levels (<1.5, 1.5~3.0, 3.0~4.5, 4.5~6, >6). The five ranks for semantic similarity and normalized shared collocates number are 0.0~0.2, 0.2~0.4, 0.4~0.6, 0.6~0.8 and 0.8 ~1.0. For every feature, we obtain a predicting value for each level after the training process. The predicting value is shown as $\frac{P(f \left| S_c \right.)}{P(f)}$. In line with that, *P(MI>6)* means the probability of all VN collocations retrieved from BNC in which the MI value is higher than 6 whereas $P(MI>6 \left| S_c \right.)$ shows the probability of all correct VN collocations with the MI value higher than 6.

Different combinations of the three features are made on the basis of the probabilistic model described in Section 4.4. Seven models derive from such combinations (See Table 1). Table 2 shows the precision of k-best suggestions for each model.

Models	Feature(s) considered
M 1	MI (Mutual Information)
M 2	SS (Semantic Similarity)
M 3	SC (Shared Collocates)
M 4	MI + SS
M 5	MI + SC
M 6	SS + SC
M 7	MI + SS + SC

Table 1. Models of feature combinations.

K-Best	M1	M2	M3	M4	M5	M6	M7
1	16.67	40.48	22.62	48.81	29.76	55.95	53.57
2	36.90	53.57	38.10	60.71	44.05	63.1	67.86
3	47.62	64.29	50.00	71.43	59.52	77.38	78.57
4	52.38	67.86	63.10	77.38	72.62	80.95	82.14
5	64.29	75.00	72.62	83.33	78.57	83.33	85.71
6	65.48	77.38	75.00	85.71	83.33	84.52	88.10
7	67.86	80.95	77.38	86.90	86.90	86.9	89.29
8	70.24	83.33	82.14	86.90	89.29	88.1	91.67
9	72.62	86.90	85.71	88.10	92.86	90.48	92.86
10	76.19	86.90	88.10	88.10	94.05	90.48	94.05

Table 2. The precision rate of Model 1- 7.

K-Best	M2	M6	M7
1	aim	*obtain	*acquire
2	generate	share	share
3	draw	*develop	*obtain
4	*obtain	generate	*develop
5	*develop	*acquire	*gain

Table 3. The K-Best suggestions for *get knowledge.*

Table 2 shows that, considering the results for each feature run separately (M1-M3), the feature 'semantic similarity' (M2) outperforms the other two. Among combined feature models (M4-M7), M7 (MI + SS+ SC), provides the highest proportion of true positives at every value of k except k = 1. The full hybrid of all three features (M7) outperforms any single feature. The best results are achieved when taking into account both statistical and semantic features. This is illustrated with results for the example *get knowledge* in Table 3 (the asterisks (*) indicate the true positives.)

6 Conclusion

In this report of work in progress, we present a probabilistic model that adopts word association measurement, semantic similarity and shared collocates in looking for corrections for learners' miscollocations. Although only VN miscollocations are examined, the model is designed to be applicable to other types of miscollocations. Applying such mechanisms to other types of miscollocations as well as detecting miscollocations will be the next steps of this research. Further, a larger amount of miscollocations should be included in order to verify our approach and to address the issue of the small drop of the full-hybrid M7 at k=1.

Acknowledgments

The work reported in this paper was partially supported by the grants from the National Science Council, Taiwan (Project Nos. 96-2524-S-008-003- and 98-2511-S-008-002-MY2)

References

Anne. Li-E Liu 2002. *A Corpus-based Lexical Semantic Investigation of VN Miscollocations in Taiwan Learners' English*. Master Thesis, Tamkang University, Taiwan.

Anthony P Cowie and Peter Howarth. 1995. Phraseological Competence and Written Proficiency, Paper Presented at the *British Association of Applied Linguistics Conference* (BAAL), Southampton, England, September.

Christina Gitsaki, Nagoya Shoka Daigaku, and Richard P. Taylor. 2000. English Collocations and Their Place in the EFL Classroom, Available at: http://www.hum.nagoya-cu.ac.jp/~taylor/publications/collocations.html.

Claudia Leacock and Martin Chodorow. 2003. Automated Grammatical Error Detection, In MD Shermis & JC Burstein (Eds.), *Automated Essay Scoring: A Cross-disciplinary*, Mahwah, NJ: Lawrence Erlbaum Associates.

David Wible, Chin-Hwa Kuo, Nai-Lung Tsao, Anne Li-E Liu, and Hsiu-Lin Lin. 2003. Bootstrapping in a Language Learning Environment. *Journal of Computer Assisted Learning*, 19, 90-102.

George Miller. 1990. WordNet: An On-line Lexical Database, *International Journal of Lexicography.*

Kenji Kita and Hiroaki Ogata. 1997. Collocations in Language Learning: Corpus-based Automatic compilation of Collocations and Bilingual Collocation Concordancer, *Computer Assisted Language Learning*. Vol.10, No. 3, 229-238.

Kenneth Church, William Gale, Patrick Hanks and Donald Hindle. 1991. Using Statistics in Lexical Analysis, in Zernik (ed), *Lexical Acquisition: Using On-line Resources to Build a Lexicon, Lawrence Erlbaum*, pp. 115-164.

Martin Chodorow, Joel R. Tetreault and Na-Rae Han. 2007. Detection of Grammatical Errors Involving Prepositions, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Special Interest Group on Semantics*, Workshop on Prepositions, 25-30.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko, Lucy Vanderwende. 2008. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction, *Proceedings of The Third International Joint Conference on Natural Language Processing*, Hyderabad, India.

Na-Rae Han, Martin Chodorow and Claudia Leacock. 2004. Detecting Errors in English Article Usage with a Maximum Entropy Classifier Trained on a Large, Diverse Corpus, *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Nai-Lung Tsao, David Wible and Chin-Hwa Kuo. 2003. Feature Expansion for Word Sense Disambiguation, *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, 126-131.

Peter Howarth. 1998. Phraseology and Second Language Acquisition. *Applied Linguistics*. 19/1, 24-44.

Yoko Futagi, Paul Deane, Martin Chodorow & Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*,21:4,353 — 367