Automatic Identification of English Collocation Errors based on Dependency Relations

Zhao-Ming Gao
National Taiwan University
zmgao@ntu.edu.tw

Abstract

We present an English miscollocation identification system based on dependency relations drawn from the Stanford parser. We test our system against a subset of error-tagged Chinese Learner English Corpus (CLEC)and obtain an overall precision of 0.75. We describe some applications and limitations of our system and suggest directions for future research.

1. Introduction

Collocations play a very important role in second language learning (cf. Lewis, 1993). They reflect users'depth of vocabulary knowledge as well as their language proficiency levels (cf.Schimitt, 2000, 2010; Nation, 2001; Nation and Webb, 2011).Researchhas shown that collocations are one of the most significant feature which distinguishes native from non-native writings. Furthermore, non-native writers tend to makecollocation errors unconsciously, many of which arise from first language interference. All these suggest the necessity of developing a miscollocation identification system to help learners detect their collocation errors as well as raise their language awareness. .Such a system might also havegreat impact for second language acquisition (SLA) research, as collections and analyses of collocation errors are vital to our understanding of the difficulties and problems learners encounter (cf. Nesselhauf, 2005). Just like other errors in learner corpora, error-tagged miscollocations are not widely and readily accessible to researchers. Traditionally, miscollocations can only be identified viavery time-consuming process of manual error tagging. Thanks to recent advance in natural language processing (NLP), automatic identification of miscollocations has been made possible. This paper presents an English miscollocation identification system by drawing on NLP tools and resources such as the Stanford parser, Google 1T ngrams, and WordNet. We will show that such a system not only has pedagogical valuebut also can facilitate the study of English miscollocations by non-native speakers.

2. Literature Review

There are two approaches to the study of collocations, namely, the frequency-based approach (Sinclair, 1987) and the phraseological approach (Cowie, 1981; Benson, 1989). Drawing on natural language processing tools, researchers have proposed automated procedures to retrieve collocations from corpora by using statistical methods such as mutual information and t-score (Church and Hanks, 1990) as well as log likelihood ratio (Dunning , 1993).In addition to statistical measures, dependency relations derived from parsers play an important role in identifying collocations (cf. Church and Hanks, 1990; Smadja, 1993;Kilgarriff, 2004).

(Jian, Chang, and Chang, 2003) present TANGO, a program which given a keyword and its part-of-speech can extract English example of four English collocation patterns (i.e. v-n, n-p, v-n-p, a-n) together with their Chinese translations from parallel corpora.

(Shei and Pain, 2000) present a conceptual frameworkto detect and correct collocation errors by Chinese learners of English. They draw on a learner corpus, a reference corpus, a dictionary of synonyms derived from WordNet, and a paraphrase database compiled using learner data. Addressing the same problem of miscollocations caused by first language interference, (Chang et al., 2008) focus on the identification and correction of V-N miscollocations by Chinese learners of English. They extract V-N collocations from British National Corpus (BNC) and leaner corpora and use a bilingual English-Chinese dictionary to identify the meanings intended by the learners. They then use the collocations extracted from BNC to pinpoint the miscollocations in the learner corpora and suggest correct collocations which learners intended to use. (Futagi et al. 2008)notice that some collocation errors are in fact due to spelling errors. They use spelling checkers to identify and correct misspelled words. They then identify miscollocation candidates by part-of-speech tags and rank-ratio statistics calculated over 1 billion word corpus by native speakers.

3. Using Dependency Relations to Identify Collocations

We follow the phraseological approach taken by(Cowie, 1981; Benson, 1989) and consider collocations a type of word combinations. As pointed out by Smadja (1993), many collocations involve pedicative relations such as subject-verb, verb-object, adjective-noun. These word combinations are easier to identify by using dependency parsers than statistical measures such as mutual information and t-score, which are useful to finding significant collocations and idioms. Our proposed miscollocation identification system is based on authentic English corpora of 14.5 million words. The system follows the lines of (Church, 1990; Smadja, 1993, Lin, 1998; Kilgarriff, 2004) in using parsers to retrieve collocations. Our approach consists of three major steps. The first step is to identify and correct spelling errors. The second step is.to identify and store the predicative relations (also known as dependency relations) occurring in the reference corpus in a dependency relation database The third step is to identify the dependency relations in a learner sentence and check them against the database of dependency relations derived from reference corpus. The technology underlying the system is similar to (Lin, 1998; Kilgarriff, 2004).

To identify dependency relations in an English sentence, the Stanford parser is used (c.f.de Marneffe, 2006). Stanford parser can identify numerous dependency relations, including modifier-noun, subject-verb, verb-noun, etc. (1) is the output of the Stanford parser, which outputs the part-of-speech tags of each word in the sentence, its syntactic structures, and dependency relations. For example, the relationnn (prices-2, Stock-1) in (1)indicates that the first word ‘Stock’ modifies the second word ‘prices’ and form a N-N dependency relation. Similarly, the second word ‘prices’ and the third word ‘plunged’ form a subject-verb relation.

(1) Stock prices plunged on many global markets Monday.

```
    Stock/NNP prices/NNS plunged/VBD on/IN
    many/JJ global/JJ markets/NNS
    Monday/NNP

(ROOT
  (S
    (NP (NNP Stock) (NNS prices))
    (VP (VBD plunged)
      (PP (IN on)
        (NP (JJ many) (JJ global) (NNS
markets)))
      (NP (NNP Monday)))))

nn(prices-2, Stock-1)
nsubj(plunged-3, prices-2)
prep(plunged-3, on-4)
amod(markets-7, many-5)
amod(markets-7, global-6)
pobj(on-4, markets-7)
dobj(plunged-3, Monday-8)
```

The performance of the Stanford parser varies with the complexity of the input sentence. If the sentence is short and the structure is not ambiguous or complicated, it can achieve relatively high accuracy.

There are six major types of dependency relations stored in our database, namely, subject-verb, verb-object, verb-adverb, noun-noun, adjective-noun, and adverb-adjective.

We use two corpora. The first is a reference corpus totaling 14.5 million words extracted from authentic English texts (i.e. the reference corpus). The second is an error-tagged learner corpus used to evaluate the accuracy of our system. The learner corpus is the subcorpus st2 in the Chinese Learner English Corpus (CLEC) and totals 251558 tokens. Each sentence in the reference corpus has been parsed by the Stanford parser to extract the dependency relations. Important dependency relations such as subject-verb, verb-object, adjective-noun, verb-adverb, and noun-noun are identified and stored in the dependency relation database for the reference corpus. The tables of.dependency relation database include the information of ahead word (the primary key in the database), its part-of-speech, the dependency relation between the headword and its collocation, the collocate of the headword, as well as the part-of-speech of the collocate. The part-of-speech information of the keyword includes noun, verb, adjective, adverb, and preposition. Nouns in the subject and object positions are distinguished to facilitate the retrieval of subject-verb and verb-object relation. Preposition is included for collocational patterns involving a verb and a preposition (e.g. ‘rely on’) or a noun and a preposition (e.g. ‘under attack’).

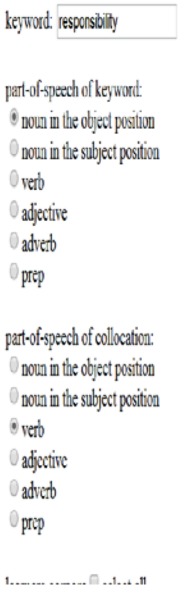
4. Identifying Miscollocations

A program is written which converts the dependency relation database into a collocation database. When a query is made, the program will search the collocation database, find all the collocations of the word in accordance with the conditions input by the user. Figure 1 is the interface of the collocation retrieval system. If the user inputs the keyword “responsibility”, “noun in the object position’ as its part-of-speech, and “verb” as the part-of-speech of the collocate, the system will return a list of potential verb collocates of the noun ‘responsibility’ such as: ‘take’, ‘shoulder’, ‘fulfill’, ‘bear’, ‘assume’, ‘accept’, ‘have’, ‘evade’, ‘shirk’. ‘avoid’.

It should be noted that the frequency information and the dependency relations we use in our program are based on lemmas (i.e. the basic form of a word). For instance, *take*, *took*, *taken*, *taking*, *takes* all have the same lemma ‘take’. We use WordNet 3.1 for converting a word into its lemma.

Following (Futagi, 2008), we identify and correct spelling errors in learner sentences in order to identify more miscollocations. We incorporate the open source spelling checker A spell and the information of language model based on the Google 1T ngram data. The correct spelling is chosen if the candidate word is the closest to the wrongly spelled word in terms of minimal edit distance and ngram probabilities.

Figure 1. The Inteface of our collocation retrieval system



keyword: responsibility

part-of-speech of keyword:
● noun in the object position
○ noun in the subject position
○ verb
○ adjective
○ adverb
○ prep

part-of-speech of collocation:
○ noun in the object position
○ noun in the subject position
● verb
○ adjective
○ adverb
○ prep

learner corpus: □ select all

Each of the dependency relations extracted from learners’ sentences not involving a personal pronoun or a proper name is checked against our English collocation retrieval program. Personal pronouns and proper names are identified by using the part-of-speech tag information output by the Stanford parser. Dependency relations with these tags are directly ignored by our collocation checker.

If a dependency relation in a learner sentence cannot be found in our English collocation database, it is considered a candidate of miscollocation.

5. Evaluations

We test our proposed system usingst2, a 251558 token subcorpus of the Chinese Learner English Corpus (CLEC)(cf. Gui and Yang, 2003), whose error tags facilitate automatic evaluation of our system. There are six types of collocation errors in the CLEC including CC1 (noun-noun), CC2 (noun-verb), CC3 (verb-noun), CC4 (adjective-noun), CC5 (verb-adverb), and CC6 (adverb-adjective). The precision rates of the six types of collocation errors are 0.77, 0.87, 0.72, 0.75, 0.83, and 0.63, respectively. Our system performs the best with CC2 (noun-verb), which has0.87 accuracy. The lowest precision is 0.63 found in CC6 (Adverb Adjective). The overall precision rate is about 0.75.

Table 1. Precision of our proposed method

	CC1 NN	CC2 NV	CC3 VN	CC4 AN	CC5 V adv	CC6 Adv A
precision	0.77	0.87	0.72	0.75	0.83	0.63

The recall rateis much lower than the precision rate, suggesting that there are many miscollocations that cannot be identified by our program.

Our dependency-based collocation extraction program has a number of limitations. As with the other collocation extraction programs, our program is not entirely reliable. Our approach fails (1) when the parser does not derive the correct dependency relations (2) or when the collocation does not belong to any dependency relation in the Stanford Parser (3)or when certain correct collocations do not occur in the reference corpus. Incorrect analyses of dependency relations typically result from sentences which have ellipsis or complicated structures. Some errors in the dependency relations are caused by the incorrect identification of the head noun in a noun phrase. One major problem with our system is the relatively small size of our reference corpus, which hasonly 14.5 million words. Another problem of using dependency relations to identify miscollocations arises from the multiple meanings and constructions a word might be associated with. Consider the word combinations of 'make stomach' and 'take university'. At first sight, they seem odd. However, inspection of the examples in (2) suggest that these word combinations are appropriate in the following contexts.

(2) (a). So I devour those buns and noodle and this fast movement of mouth makes my stomach uncomfortable a whole morning.
(b). TakeNational Don HuaUniversityfor instance.

In other words, using dependency relations to identify miscollocations might be inadequate when the keyword in question has different meanings and can appear in different constructions. This is a serious limitation to the dependency-based approach to miscollocation identification. Solution to this problem might require identification of different constructions a word can occur in. This, however, cannot be easily achieved at present. Another limitation to our approach is that a phrase may be inappropriate even if all its parts seem acceptable, because the correctness of all the smaller parts of the phrase cannot entail the correctness of the larger units. The same applies to ngrams and dependency relations. Just like ngrams, dependency relations are approximations to larger units such as a phrase or a sentence. They alone cannot give us all the information about their grammatical status or contextual appropriateness of which they are a part.

6. Applications

One of the applications of our program is automatic identification of collocational differences in learner and authentic corpora. With this function, we are able to automatically collect miscollocationsfrom learner corpora. For example, by inputting the verb 'take' and the part-of-speech of a noun in the object position, we extract 'take exercise', 'take adventure', 'take reform', 'take lecture', 'take grade', and 'take travel'asmiscollocations.'Some examples containing thesemiscollocations in the learner corpora are listed in (3).

3. (a). Theytakemore**exercises**than ever.
(b). They like new things and like taking **adventure**.
(c). We take the **reform** and open policy.
(d).I have to take the economic **lectures** and learn to use computer in order to gain more knowledge and keep up with the society.
(e). In junior high school, the English teacher only taught you how to take good **grade** in the test.

Some other examples of miscollocations identified by our system are provided in (4).

(4) (a). That will open our sights of the world.
(b). Since we have faced the cricis of fresh water, we should do what we can to release the **problem**.
(c). Meanwhile, on the way to Belcy I planned to **take**a travel in the famous cities.
(d). What defines a really alive **person** is his personal functions but not physiological ones.
(e). Nonprofit organizations do many **efforts** to the world.
(f). We not only learn the knowledge of financial management but also make **action** for it.
(h). It has long been a controversy that a teacher should take physical **punishment** or education by love to teach their students.

With our system, it is relatively easy to find general patterns about learners' miscollocations. First, learners have difficulties in collocations involving support verbs, such as 'take', 'make', and 'do'. They are often confused about which support verb they should use in a certain context (cf. (3a)-(3e), (4e)-(4h)).Second, learners are heavily influenced by their first language and cannot distinguish the subtle nuances between near synonyms (e.g. 'widen' or 'broaden' vs. 'open', 'vision' vs. 'sight' in (4a), 'trip' vs. 'travel' in (4c), and 'living' vs. 'alive' in (4d)). Third, learners are not only confused by semantically similar words but also phonetically or orthographically similar words (e.g. 'relieve' vs. 'release' in (4b)).

The examples in (3) and (4) show that our systems can efficiently and effectively identify common miscollocation patterns and facilitate research in L2 miscollocations in a way similar to (Nesselhauf, 2005). Clearly, our system ismuch more efficient than traditional method of manual error tagging in identifying miscollocations as well as differences between native and non-native usage.

7. Conclusions and Future Research

The proposed English miscollocation checker might help learners reduce collocation errors and develop learner autonomy. It has the potential of alleviating teachers' burden in correcting students' English miscollocations.The proposed system can automatically collect and characterize the collocational differences used in learner and authentic corpora. Thisfeature might have positive impact for the teaching, learning, and research of collocations and miscollocations.

There are a number of limitations to our approach. For example, the corpus size of our reference corpus is not large enough. The accuracy of the dependency relations derived from the Stanford parser should also be improved. There are also constructions which cannot be adequately analyzed by dependency relations. These constructions allow greater flexibility than dependency relations.

While our proposed system for identifying collocation errors are not completely reliable, they might help learners improve their writing if the tool is used properly. Future research includes (1). qualitative and quantitative evidence of the learning effects of the proposed system in second language writing (2).development of an intelligent system that can not only detect but also correct collocation errors (3). investigation of the relationships between miscollocation types, error gravity, and learners' proficiency levels.

References

Benson, M. 1989. The Structure of the Collocational Dictionary."International Journal of Lexicography, Vol. 2, No. 1, pp. 1-14.

Cowie, A. P. 1981. The Treatment of Collocations and Idioms in Learner's Dictionaries. Applied Linguistics, Vol. 2, No, 3, pp. 223-235.

Chang, Y.-C., Chang Jason, Chen Hao-Jan, & Liou, H.C. 2008. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. Computer Assisted Language Learning, 21(3), 283-299.

Church, Ken. and Hanks, Patrick. 1990 Word Association Norms, Mutual Information, and Lexicography." Computational Linguistics, Vol. 16, No. 1, pp. 22-29.

Cowie, Anthony. 1981. The Treatment of Collocations and Idioms in Learner's Dictionaries.Applied Linguistics, Vol. 2, No, 3, pp. 223-235.

Chang, Yu.-Chia., Chang, Jason, Chen Hao-Jan, & Liou, Hsien-Chin. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learnin*g, 21(3), 283-299.

deMarneffe, Marie-Catherine, MacCartney, Bill and Manning, Christopher. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Futagi, Yoko, et al. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. Computer-Assisted Language Learning, Vol. 21, No. 4, pp.353 – 367.

Gui, Shicun and Yang, Huizhong. 2002. Chinese Learner English Corpus. Foreign Language Education Press, Shanghai.

Jian, J.-Y., Chang, Y.-C., Chang, J.-S. 2004. Tango: Bilingual Collocational Concordancer. Poster presented at the Annual Conference of the Association for Computational Linguistics.

Kilgarriff, Adam et al. 2004. The Sketch Engine.In Proceedings of EURALEX, Lorient, France.

Lewis, Michalel. 1993. The Lexical Approach: the State of ELT and a Way Forward.: Thompson/Heinle, Boston.

Lin, Dekang. 1998. Extracting Collocations from Text Corpora. First Workshop on Computational Terminology, Montreal, Canada, August, 1998.

Nation, Paul. 2001. Learning Vocabulary in Another Language. Cambridge: University Press, Cambridge.

Nation, Paul, and Webb, Stuart. 2011. Researching and Analyzing Vocabulary.Heinle, Boston

Nesselhauf, N. 2005.Collocations in a Learner Corpus. John Benjaimins. Amsterdam.

Schmitt, Norbert. 2000. Vocabulary in Language Learning.

Schimitt,Nobert. 2010. Researching Vocabulary: a Vocabulary Research Manual. Palgrave Macmillian, London.

Smadja, Frank. 1993Retrieving Collocations from Text: Xtract. ComputationalLinguistics, Vol. 19, No. 1, pp. 143 - 177.

Shei, Chi.-Chiang.and Pain, Helen. 2000. An ESL Writer's Collocation Aid Computer-Assisted Language Learning, Vol. 13, No. 2, pp. 167-182.

Sinclair, John. 1987. Collocations: a Progress Report. In Steele and Thrreadgold (eds.) , 1987, Language Topics: Essays in Honour of Michael Halliday. John Benjamins, Amsterdam and Philadelphia.

Software Used in this Study

Aspellhttp://aspell.net/

Chinese Learner English Corpus. CD accompanying (Gui and Yang, 2003)

Google 1T ngramshttp://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

The Stanford Parser.http://webdocs.cs.ualberta.ca/~lindek/Stanford.htm

WordNet 3.1http://wordnet.princeton.edu/