Correcting Semantic Collocation Errors with L1-induced Paraphrases

Daniel Dahlmeier[1] and **Hwee Tou Ng**[1,2]
[1]NUS Graduate School for Integrative Sciences and Engineering
[2]Department of Computer Science, National University of Singapore
{danielhe,nght}@comp.nus.edu.sg

Abstract

We present a novel approach for automatic collocation error correction in learner English which is based on paraphrases extracted from parallel corpora. Our key assumption is that collocation errors are often caused by semantic similarity in the first language (L1-language) of the writer. An analysis of a large corpus of annotated learner English confirms this assumption. We evaluate our approach on real-world learner data and show that L1-induced paraphrases outperform traditional approaches based on edit distance, homophones, and WordNet synonyms.

1 Introduction

Grammatical error correction (GEC) is emerging as a commercially attractive application of natural language processing (NLP) for the booming market of English as foreign or second language (EFL/ESL[1]).

The de facto standard approach to GEC is to build a statistical model that can choose the most likely correction from a *confusion set* of possible correction choices. The way the confusion set is defined depends on the type of error. Work in context-sensitive spelling error correction (Golding and Roth, 1999) has traditionally focused on confusion sets with similar spelling (e.g., {*dessert*, *desert*}) or similar pronunciation (e.g., {*there*, *their*}). In other words, the words in a confusion set are deemed confusable because of orthographic or phonetic similarity. Other work in GEC has defined the confusion sets based on syntactic similarity, for example all English articles or the most frequent English prepositions form a confusion set (see for example (Tetreault et al., 2010; Rozovskaya and Roth, 2010; Gamon, 2010; Dahlmeier and Ng, 2011) among others).

In contrast, we investigate in this paper a class of grammatical errors where the source of confusion is the similar *semantics* of the words, rather than orthography, phonetics, or syntax. In particular, we focus on *collocation errors* in EFL writing. The term *collocation* (Firth, 1957) describes a sequence of words that is conventionally used together in a particular way by native speakers and appears more often together than one would expect by chance. The correct use of collocations is a major difficulty for EFL students (Farghal and Obiedat, 1995).

In this work, we present a novel approach for automatic correction of collocation errors in EFL writing. Our key observation is that words are potentially confusable for an EFL student if they have similar translations in the writer's first language (L1-language), or in other words if they have the same semantics in the L1-language of the writer. The Chinese word 看 (*kàn*), for example, has over a dozen translations in English, including the words *see*, *look*, *read*, and *watch*. A Chinese speaker who still "thinks" in Chinese has to choose from all these possible translations when he wants to express a sentence like *I like to **watch** movies* and might instead produce a sentence like **I like to **look** movies*. Although the meanings of *watch* and *look* are similar, the former is clearly the more fluent choice in this context. While these types of *L1-transfer er-*

[1]For simplicity, we will collectively refer to both terms as *English as a foreign language (EFL)*

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 107–117,
Edinburgh, Scotland, UK, July 27–31, 2011.

rors have been known in the EFL teaching literature (Swan and Smith, 2001; Meng, 2008), research in GEC has mostly ignored this fact.

We first analyze collocation errors in the NUS Corpus of Learner English (NUCLE), a fully annotated one-million-word corpus of learner English which we will make available to the community for research purposes (see Section 3 for details about the corpus). Our analysis confirms that many collocation errors can be traced to similar translations in the writer's L1-language. Based on this result, we propose a novel approach for automatic collocation error correction. The key component in our approach generates *L1-induced paraphrases* which we automatically extract from an L1-English parallel corpus. Our proposed approach outperforms traditional approaches based on edit distance, homophones, and WordNet synonyms on a test set of real-world learner data in an automatic and a human evaluation. Finally, we present a detailed analysis of unsolved instances in our data set to highlight directions for future work.

Our work adds to a growing body of research that leverages parallel corpora for semantic NLP tasks, for example in word sense disambiguation (Ng et al., 2003; Chan and Ng, 2005; Ng and Chan, 2007; Zhong and Ng, 2009), paraphrasing (Bannard and Callison-Burch, 2005; Liu et al., 2010a), and machine translation evaluation (Snover et al., 2009; Liu et al., 2010b).

The remainder of this paper is organized as follows. The next section reviews related work. Section 3 presents our analysis of collocation errors. Section 4 describes our approach for automatic collocation error correction. The experimental setup and the results are described in Sections 5 and 6, respectively. Section 7 provides further analysis. Section 8 concludes the paper.

2 Related Work

In this section, we give an overview of related work on collocation error correction. We also highlight differences between collocation error correction and related NLP tasks like context-sensitive spelling error correction, synonym extraction, lexical substitution, and paraphrasing.

Most work in collocation error correction has relied on dictionaries or manually created databases to generate collocation candidates (Shei and Pain, 2000; Wible et al., 2003; Futagi et al., 2008). Other work has focused on finding candidates that collocate with similar words, e.g., verbs that appear with the same noun objects form a confusion set (Liu et al., 2009; Wu et al., 2010). The work most similar to ours is probably the one presented by Chang *et al.* (2008), as they also use translation information to generate collocation candidates. However, they do not use automatically derived paraphrases from parallel corpora but bilingual dictionaries. Dictionaries usually have lower coverage, do not contain longer phrases or inflected forms, and do not provide any translation probability estimates. Also, their work focuses solely on verb-noun collocations, while we target collocations of arbitrary syntactic type.

Context-sensitive spelling error correction is the task of correcting spelling mistakes that result in another valid word, see for example (Golding and Roth, 1999). It has traditionally focused on a small number of pre-defined confusion sets, like homophones or frequent spelling errors. Even when the confusion sets are formed automatically, the similarity of words in a confusion set has been based on edit distance or phonetic similarity (Carlson et al., 2001). In contrast, we focus on words that are confusable due to their similar semantics instead of similar spelling or pronunciation. Also, we do not assume that the set of confusion sets is already given to us. Instead, we automatically extract confusable candidates from a parallel corpus.

Synonym extraction (Wu and Zhou, 2003), lexical substitution (McCarthy and Navigli, 2007) and paraphrasing (Madnani and Dorr, 2010) are related to collocation correction in the sense that they try to find semantically equivalent words or phrases. However, there is a subtle but important difference between these tasks and collocation correction. In the former, the main criterion is whether the original phrase and the synonym/paraphrase candidate are substitutable, i.e., both form a grammatical sentence when substituted for each other in a particular context. In contrast, in collocation correction, we are primarily interested in finding candidates which are *not substitutable* in their English context but *appear to be substitutable* in the L1-language of the writer, i.e., one forms a grammatical English sentence but the other does not.

Sentences	52,149
Words	1,149,100
Distinct words	27,593
Avg. sentence length (words)	22.04
Collocation errors	2,747
Avg. collocation error length (words)	1.17
Avg. correction length (words)	1.13

Table 1: Statistics of the NUS Corpus of Learner English (NUCLE)

3 Analysis of EFL collocation errors

While the fact that collocation errors can be caused by L1-transfer has been ascertained by EFL researchers (Meng, 2008), we need to quantify how frequent collocation errors can be traced to these types of transfer errors in order to estimate how many errors in EFL writing we can potentially hope to correct with information about the writer's L1-language.

We base our analysis on the NUS Corpus of Learner English (NUCLE). The corpus consists of about 1,400 essays written by EFL university students on a wide range of topics, like environmental pollution or healthcare. Most of the students are native Chinese speakers. The corpus contains over one million words which are completely annotated with error tags and corrections. All annotations have been performed by professional English instructors. The statistics of the corpus are summarized in Table 1. The annotation is stored in a stand-off fashion. Each error tag consists of the start and end offset of the annotation, the type of the error, and the appropriate gold correction as deemed by the annotator. The annotators were asked to provide a correction that would result in a grammatical sentence if the selected word or phrase would be replaced by the correction.

In this work, we focus on errors which have been marked with the error tag *wrong collocation/idiom/preposition*. As preposition errors are not the focus of this work, we automatically filter out all instances which represent simple substitutions of prepositions, using a fixed list of frequent English prepositions. In a similar way, we filter out a small number of article errors which were marked as collocation errors. Finally, we filter out instances where the annotated phrase or the suggested correction is longer than 3 words, as we observe that they contain highly context-specific corrections and are unlikely to generalize well (e.g., *"for the simple reasons that these can help them"* $\rightarrow$ *"simply to"*).

After filtering, we end up with 2,747 collocation errors and their respective corrections, which account for about 6% of all errors in NUCLE. This makes collocation errors the 7th largest class of errors in the corpus after article errors, redundancies, prepositions, noun number, verb tense, and mechanics. Not counting duplicates, there are 2,412 distinct collocation errors and corrections. Although there are other error types which are more frequent, collocation errors represent a particular challenge as the possible corrections are not restricted to a closed set of choices and they are directly related to *semantics* rather than syntax. We analyzed the collocation errors and found that they can be attributed to the following sources of confusion:

Spelling: We suspect that an error is caused by similar orthography if the edit distance between the erroneous phrase and its correction is less than a certain threshold.

Homophones: We suspect that an error is caused by similar pronunciation if the erroneous word and its correction have the same pronunciation. We use the CuVPlus English dictionary (Mitton, 1992) to map words to their phonetic representations.

Synonyms: We suspect that an error is caused by synonymy if the erroneous word and its correction are synonyms in WordNet (Fellbaum, 1998). We use WordNet 3.0.

L1-transfer: We suspect that an error is caused by L1-transfer if the erroneous phrase and its correction share a common translation in a Chinese-English phrase table. The details of the phrase table construction are described in Section 4. We note that although we focus on Chinese-English translation, our method is applicable to any language pair where parallel corpora are available.

As CuVPlus and WordNet are defined for individual words, we extend the matching process to phrases in the following way: two phrases A and B are deemed homophones/synonyms if they have the same length and the i-th word in phrase A is a homophone/synonym of the corresponding i-th word in phrase B.

Spelling	... it received *critics (criticism)* as much as complaints ...
	... budget for the aged to *improvise (improve)* other areas.
Homophones	... diverse spending can *aide (aid)* our country.
	... *insure (ensure)* the safety of civilians ...
Synonyms	... rapid *increment (increase)* of the seniors ...
	... energy that we can *apply (use)* in the future ...
L1-transfer	... and *give (provide,* 给予 *)* reasonable fares to the public ...
	... and *concerns (attention,* 关注 *)* that the nation put on technology and engineering ...

Table 3: Examples of collocation errors with different sources of confusion. The correction is shown in parenthesis. For L1-transfer, we also show the shared Chinese translation. The L1-transfer examples shown here do not belong to any of the other categories.

Suspected Error Source	Tokens	Types
Spelling	154	131
Homophones	2	2
Synonyms	74	60
L1-transfer	1016	782
L1-transfer w/o spelling	954	727
L1-transfer w/o homophones	1015	781
L1-transfer w/o synonyms	958	737
L1-transfer w/o spelling, homophones, synonyms	906	692

Table 2: Analysis of collocation errors. The threshold for spelling errors is one for phrases of up to six characters and two for the remaining phrases.

The results of the analysis are shown in Table 2. Tokens refer to running erroneous phrase-correction pairs including duplicates, and types refer to distinct erroneous phrase-correction pairs. As a collocation error can be part of more than one category, the rows in the table do not sum up to the total number of errors. The number of errors that can be traced to L1-transfer greatly outnumbers all other categories. The table also shows the number of collocation errors that can be traced to L1-transfer but not the other sources. 906 collocation errors with 692 distinct collocation error types can be attributed only to L1-transfer but not to spelling, homophones, or synonyms. Table 3 shows some examples of collocation errors for each category from our corpus. We note that there are also collocation error types that cannot be traced to any of the above sources. We will return to these errors in Section 7.

4 Correcting Collocation Errors

In this section, we propose a novel approach for correcting collocation errors in EFL writing.

4.1 L1-induced Paraphrases

We use the popular technique of paraphrasing with parallel corpora (Bannard and Callison-Burch, 2005) to automatically find collocation candidates from a sentence-aligned L1-English parallel corpus. As most of the essays in our corpus are written by native Chinese speakers, we use the FBIS Chinese-English corpus, which consists of about 230,000 Chinese sentences (8.5 million words) from news articles, each with a single English translation. We tokenize and lowercase the English half of the corpus in the standard way. We segment the Chinese half of the corpus using the maximum entropy segmenter from (Ng and Low, 2004; Low et al., 2005). Subsequently, we automatically align the texts at the word level using the Berkeley aligner (Liang et al., 2006; Haghighi et al., 2009). We extract English-L1 and L1-English phrases of up to three words from the aligned texts using the widely used phrase extraction heuristic in (Koehn et al., 2003). The paraphrase probability of an English phrase e_1 given an English phrase e_2 is defined as

$$p(e_1|e_2) = \sum_f p(e_1|f)p(f|e_2) \quad (1)$$

where f denotes a foreign phrase in the L1 language. The phrase translation probabilities $p(e_1|f)$ and $p(f|e_2)$ are estimated by maximum likelihood estimation and smoothed using Good-Turing smoothing (Foster et al., 2006). Finally, we only keep para-

phrases with a probability above a certain threshold (set to 0.001 in our work).

4.2 Collocation Correction with Phrase-based SMT

We implement our approach in the framework of phrase-based statistical machine translation (SMT) (Koehn et al., 2003). Phrase-based SMT tries to find the highest scoring translation $\mathbf{e}$ given an input sentence $\mathbf{f}$. The *decoding* process of finding the highest scoring translation is guided by a log-linear model which scores translation candidates using a set of feature functions $h_i,\ i = 1, \ldots, n$

$$score(\mathbf{e}|\mathbf{f}) = \exp\left(\sum_{i=1}^{n} \lambda_i h_i(\mathbf{e}, \mathbf{f})\right) . \quad (2)$$

Typical features include a phrase translation probability $p(e|f)$, an inverse phrase translation probability $p(f|e)$, a language model score $p(\mathbf{e})$, and a constant phrase penalty. The optimization of the feature weights $\lambda_i,\ i = 1, \ldots, n$ can be done using minimum error rate training (MERT) (Och, 2003) on a development set of input sentences and their reference translations.

Because of the great flexibility of the log-linear model, researchers have used the framework for other tasks outside SMT, including grammatical error correction (Brockett et al., 2006). We adopt a similar approach in this work. We modify the phrase table of the popular phrase-based SMT decoder MOSES (Koehn et al., 2007) to include collocation corrections with features derived from spelling, homophones, synonyms, and L1-induced paraphrases.

- **Spelling:** For each English word, the phrase table contains entries consisting of the word itself and each word that is within a certain edit distance from the original word. Each entry has a constant feature of 1.0.

- **Homophones:** For each English word, the phrase table contains entries consisting of the word itself and each of the word's homophones. We determine homophones using the CuVPlus dictionary. Each entry has a constant feature of 1.0.

- **Synonyms:** For each English word, the phrase table contains entries consisting of the word itself and each of its synonyms in WordNet. If a word has more than one sense, we consider all its senses. Each entry has a constant feature of 1.0.

- **L1-paraphrases:** For each English phrase, the phrase table contains entries consisting of the phrase and each of its L1-derived paraphrases as described in Section 4.1. Each entry has two real-valued features: a paraphrase probability according to Equation 1 and an inverse paraphrase probability.

- **Baseline** We combine the phrase tables built for spelling, homophones, and synonyms. The combined phrase table contains three binary features for spelling, homophones, and synonyms, respectively.

- **All** We combine the phrase tables from spelling, homophones, synonyms, and L1-paraphrases. The combined phrase table contains five features: three binary features for spelling, homophones, and synonyms, and two real-valued features for the L1-paraphrase probability and inverse L1-paraphrase probability.

Additionally, each phrase table contains the standard constant phrase penalty feature. The first four tables only contain collocation candidates for individual words. We leave it to the decoder to construct corrections for longer phrases during the decoding process if necessary.

5 Experiments

In this section, we empirically evaluate our approach on real collocation errors in learner English.

5.1 Data Set

We randomly sample a development set of 770 sentences and a test set of 856 sentences from our corpus. Each sentence contains exactly one collocation error. The sampling is performed in a way that sentences from the same document cannot end up in both the development and the test set. In order to

keep conditions as realistic as possible, we make no attempt to filter the test set in any way.

We build phrase tables as described in Section 4.2. For the purpose of the experiments reported in this paper, we only need to generate phrase table entries for words and phrases which actually appear in the development or test set.

5.2 Evaluation Metrics

We conduct an automatic and a human evaluation. Our main evaluation metric is *mean reciprocal rank (MRR)* which is the arithmetic mean of the inverse ranks of the first correct answer returned by the system

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank(i)} \quad (3)$$

where N is the size of the test set. If the system did not return a correct answer for a test instance, we set $\frac{1}{rank(i)}$ to zero.

In the human evaluation, we additionally report precision at rank $k,\ k = 1, 2, 3$, which we calculate as follows:

$$\text{P@k} = \frac{\sum_{a \in A} score(a)}{|A|} \quad (4)$$

where A is the set of returned answers of rank k or less and $score(\cdot)$ is a real-valued scoring function between zero and one.

5.3 Collocation Error Experiments

Automatic correction of collocation errors can conceptually be divided into two steps: *i) identification* of wrong collocations in the input, and *ii) correction* of the identified collocations. In this work, we focus on the second step and assume that the erroneous collocation has already been identified. While this might seem like a simplification, it has been the common evaluation setup in collocation error correction (see for example (Wu et al., 2010)). It also has a practical application where the user first selects a word or phrase and the system displays possible corrections.

In our experiments, we use the start and end offset of the collocation error provided by the human annotator to identify the location of the collocation error. We fix the translation of the rest of the sentence to its identity. We remove phrase table entries where the phrase and the candidate correction are identical, thus practically forcing the system to change the identified phrase. We set the distortion limit of the decoder to zero to achieve monotone decoding. We previously observed that word order errors are virtually absent in our collocation errors. For the language model, we use a 5-gram language model trained on the English Gigaword corpus with modified Kneser-Ney smoothing. All experiments use the same language model to allow a fair comparison.

We perform MERT training with the popular BLEU metric (Papineni et al., 2002) on the development set of erroneous sentences and their corrections. As the search space is restricted to changing a single phrase per sentence, training converges relatively quickly after two or three iterations. After convergence, the model can be used to automatically correct new collocation errors.

6 Results

We evaluate the performance of the proposed method on our test set of 856 sentences, each with one collocation error. We conduct both an automatic and a human evaluation. In the automatic evaluation, the system's performance is measured by computing the rank of the gold answer provided by the human annotator in the n-best list of the system. We limit the size of the n-best list to the top 100 outputs. If the gold answer is not found in the top 100 outputs, the rank is considered to be infinity, or in other words, the inverse of the rank is zero. We also report the number of test instances for which the gold answer was ranked among the top k answers, $k = 1, 2, 3, 10, 100$. The results of the automatic evaluation are shown in Table 4

For collocation errors, there are usually more than one possible correct answer. Therefore, automatic evaluation underestimates the actual performance of the system by only considering the single gold answer as correct and all other answers as wrong. As such, we carried out a human evaluation for the systems BASELINE and ALL. We recruited two English speakers to judge a subset of 500 test sentences. For each sentence, a judge was shown the original sentence and the 3-best candidates of each of the two systems. We restricted human evaluation to the 3-best candidates, as we believe that answers at a rank

Model	Rank = 1	Rank ≤ 2	Rank ≤ 3	Rank ≤ 10	Rank ≤ 100	MRR
Spelling	35	41	42	44	44	4.51
Homophones	1	1	1	1	1	0.11
Synonyms	32	47	52	60	61	4.98
Baseline	49	68	80	93	96	7.61
L1-paraphrases	93	133	154	**216**	**243**	15.43
All	**112**	**150**	**166**	**216**	241	**17.21**

Table 4: Results of automatic evaluation. Columns two to six show the number of gold answers that are ranked within the top k answers. The last column shows the mean reciprocal rank in percentage. Bigger values are better.

P(A)	0.8076
Kappa	0.6152

Table 5: Inter-annotator agreement. $P(E) = 0.5$.

larger than three will not be very useful in a practical application. The candidates are displayed together in alphabetical order without any information about their rank or which system produced them or the gold answer by the annotator. The difference between the candidates and the original sentence is highlighted. The judges were asked to make a binary judgment for each of the candidates on whether the proposed candidate is a valid correction of the original or not. We represent valid corrections with a score of 1.0 and invalid corrections with a score of 0.0. Inter-annotator agreement is reported in Table 5. The chance of agreement $P(A)$ is the percentage of times that the annotators agree, and $P(E)$ is the expected agreement by chance, which is 0.5 in our case. The Kappa coefficient is defined as

$$\text{Kappa} = \frac{\text{P(A)} - \text{P(E)}}{1 - \text{P(E)}}$$

We obtain a Kappa coefficient of 0.6152. A Kappa coefficient between 0.6 and 0.8 is considered as showing *substantial* agreement according to Landis and Koch (1977). To compute precision at rank k, we average the judgments. Thus, a system can receive a score of 0.0 (both judgments negative), 0.5 (judges disagree), or 1.0 (both judgments positive) for each returned answer. To compute MRR, we cannot simply average the judgments as MRR requires binary judgments on whether an item is correct or not. Instead, we report MRR on the union and the intersection of the judgments. In the first case, the rank of the first correct item is the minimum rank of any item judged correct by *either* judge. In the second case, the rank of the first correct item is the minimum rank of any item judged correct by *both* judges. The results for the human evaluation are shown in Table 6. Our best system ALL outperforms the BASELINE approach on all measures. It receives a precision at rank 1 of 38.20% and a MRR of 33.16% (intersection) and 57.26% (union). Table 7 shows some examples from our test set.

Unfortunately, comparison of our results with previous work is complicated by the fact that there currently exists no standard data set for collocation error correction. We will make our corpus available for research purposes in the hope that it will allow researchers to more directly compare their results in future.

7 Analysis

In this section, we analyze and categorize those test instances for which the ALL system could not produce an acceptable correction in the top 3 candidates. We manually analyze 100 test sentences for which neither judge had deemed any candidate answer to be a valid correction. Based on our findings, we categorize the 100 sentences into eight categories which are shown below. Table 8 shows examples from each category.

Out-of-vocabulary (21/100) The most frequent reason why the system does not produce a good correction is that the erroneous collocation is out of vocabulary. These collocations often involve compound words, like *man-hours* or *carefully-nurturing*, or infrequent expressions, like *copy phenomena*, which do not appear in the FBIS parallel corpus. We expect that this problem can be reduced by using larger parallel corpora for paraphrase extraction.

Near miss (18/100) The second largest category

Model	Rank = 1	Rank $\leq$ 2	Rank $\leq$ 3	P@1	P@2	P@3	MRR
Baseline	43 | 141	69 | 201	83 | 237	18.40	16.68	15.36	12.13 | 36.60
All	**137 | 245**	**176 | 303**	**204 | 340**	**38.20**	**32.87**	**29.30**	**33.16 | 57.26**

Table 6: Results of human evaluation. Rank and MRR results are shown for the intersection (first value) and union (second value) of human judgments.

Original	it must be clear, concise and unambiguous to *prevent* any off-track
Gold	it must be clear, concise and unambiguous to *avoid* any off-track
All	it must be clear, concise and unambiguous to *avoid* any off-track
	it must be clear, concise and unambiguous to *stop* any off-track
	it must be clear, concise and unambiguous to *block* any off-track
Baseline	*it must be clear, concise and unambiguous to *present* any off-track
	it must be clear, concise and unambiguous to *forestall* any off-track
	*it must be clear, concise and unambiguous to *lock* any off-track

Original	although many may *agree* that public spending on the elderly should be limited . . .
Gold	although many may *argue* that public spending on the elderly should be limited . . .
All	although many may *believe* that public spending on the elderly should be limited . . .
	although many may *think* that public spending on the elderly should be limited . . .
	although many may *accept* that public spending on the elderly should be limited . . .
Baseline	*although many may *agreed* that public spending on the elderly should be limited . . .
	*although many may *hold* that public spending on the elderly should be limited . . .
	*although many may *agrees* that public spending on the elderly should be limited . . .

Table 7: Examples of test sentences with the top 3 answers of the ALL and BASELINE system. An answer judged incorrect by at least one judge is marked with an asterisk (*).

Out of vocabulary	. . . many illegal *copy phenomena (copy phenomena, **copies**)* in china.
	. . . lead to reduced *man-hours (man-hours, **productivity**)* as people fall sick . . .
Near miss	. . . smaller groups of people, sometimes *even (more, **only**)* individual .
	. . . take pre-emptive *actions (activities, **measures**)* . . .
Function/auxiliary words	. . . entertainment an elderly person can *have (be, **enjoy**)* .
	. . . and the security issue is solved *also (and, **too**)*
Discourse specific	. . . make other countries respect and fear *you (<question mark>, **a country**)*
	. . . will contribute nothing to the *accident (explosion, **problem**)* .
Spelling errors	this *incidence (rate, **incident**)* had also resulted in 4 fatalities . . .
	refrigerator did not *compromise (yield, **comprise**)* of any moving parts . . .
Word sense	. . . refers to the desire or shortage of a *good (better, **commodity**)* and . . .
	. . . members are always from different *majors (major league, **specialties**)*
Preposition constructions	. . . can be an area worth *investing (investing, **investing in**)*
	. . . in spending their *resources (resources, **resources on**)*
Others	this might *redirect (make sound, **reduce**)* foreign investments . . .
	. . . a trading hub since *british 's (british 's, **british**)* rule.

Table 8: Examples of sentences without valid corrections by the ALL model. The top-1 suggestion of the system and the gold answer (in bold) are shown in parenthesis.

consists of instances where the system barely misses the gold standard answer. This includes cases where the extracted L1-paraphrases do not contain the exact phrase required, e.g., the paraphrase table contains *even|||only get* when the gold correction was *even* → *only*, or the phrase table actually contains the gold answer but fails to rank it among the top 3 answers. The first problem could be addressed by modifying the phrase extraction heuristic to produce more fine-grained phrase pairs. The second problem requires a better language model. Although our language model is trained on the large English Gigaword corpus, it is not always successful in promoting the correct candidate to the top. The domain mismatch between the newswire domain of Gigaword and student essays could be one reason for this.

Function/auxiliary words (14/100) We observe that collocation errors that involve function words or auxiliary words are not handled very well by our model. Function words and auxiliary words in English lack direct counterparts in Chinese, which is why the word alignments and therefore the extracted phrases for these words contain a high amount of noise. As function words and auxiliaries are essentially a closed set, it might be more promising to build separate models with fixed confusion sets for them.

Discourse specific (14/100) Some of the gold answers are highly specific to the particular discourse that they appear in. As our model corrects collocation errors at the sentence level, such gold answers will be very difficult or impossible to determine correctly. Including more context beyond the sentence level might help to overcome this problem, although it is not easy to integrate this larger context information.

Spelling errors (9/100) Some of the collocation errors are caused by spelling mistakes, e.g., *incidence* instead of *incident*. Although the ALL model includes candidates which are created through edit distance, paraphrase candidates created from the misspelled word can dominate the top 3 ranks, e.g., *rate* and *frequently* are paraphrases of *incidence*. A possible solution would be to perform spell-checking as a separate pre-processing step prior to collocation correction.

Word sense (7/100) Some of the failures of the model can be attributed to ambiguous senses of the collocation phrase. As we do not perform word sense disambiguation in our current work, candidates from other word senses can end up as the top candidates. Including word sense disambiguation into the model might help, although accurate word sense disambiguation on noisy learner text may not be easy.

Preposition constructions (6/100) Some of the collocation errors involve preposition constructions, e.g., the student wrote *attend* instead of *attend to*. Because prepositions do not have a direct counterpart in Chinese, the L1-paraphrases do not model their semantics very well. This category is closely related to the function/auxiliary word category. Again, since prepositions are a closed set, it might be more promising to build a separate model for them.

Others (11/100) Other mistakes include collocation errors where the gold answer slightly changed the semantics of the target word, e.g., *redirect potential foreign investments* → *reduce potential foreign investments*, active-passive alternation (*enhanced economics* → *was economical*), and noun possessive errors (*british 's rule* → *british rule*).

8 Conclusion and Future Work

We have presented a novel approach for correcting collocation errors in written learner text. Our approach exploits the semantic similarity of words in the writer's L1-language based on paraphrases extracted from an L1-English parallel corpus. Our experiments on real-world learner data show that our approach outperforms traditional approaches based on edit distance, homophones, and synonyms by a large margin.

In future work, we plan to extend our system to fully automatic collocation correction that involves both identification and correction of collocation errors.

Acknowledgments

This research was done for CSIDM Project No. CSIDM-200804 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

References

C. Bannard and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*.

C. Brockett, W. B. Dolan, and M. Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of ACL*.

A. J. Carlson, J. Rosen, and D. Roth. 2001. Scaling up context-sensitive text correction. In *Proceedings of IAAI*.

Y. S. Chan and H. T. Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *Proceedings of AAAI*.

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):283–299.

D. Dahlmeier and H. T. Ng. 2011. Grammatical error correction with alternating structure optimization. In *Proceedings of ACL*.

M. Farghal and H. Obiedat. 1995. Collocations: A neglected variable in EFL. *International Review of Applied Linguistics*, 33.

C. Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge,MA.

J. R. Firth. 1957. *Papers in Linguistics 1934-1951*. Oxford University Press, London.

G. Foster, R. Kuhn, and H. Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proceedings of EMNLP*.

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. *Journal of Computer-Assisted Learning*, 21.

M. Gamon. 2010. Using mostly native data to correct errors in learners' writing: A meta-classifier approach. In *Proceedings of HLT-NAACL*.

A. R. Golding and D. Roth. 1999. A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34:107–130.

A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. 2009. Better word alignments with supervised ITG models. In *Proceedings of ACL-IJCNLP*.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL demonstration session*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

P. Liang, B. Taskar, and D. Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*.

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

C. Liu, D. Dahlmeier, and H. T. Ng. 2010a. PEM: a paraphrase evaluation metric exploiting parallel texts. In *Proceedings of EMNLP*.

C. Liu, D. Dahlmeier, and H. T. Ng. 2010b. TESLA: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of WMT and MetricsMATR*.

J. K. Low, H. T. Ng, and W. Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*.

N. Madnani and B. J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

D. McCarthy and R. Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*.

J. Meng. 2008. Erroneous collocations caused by language transfer in Chinese EFL writing. *US-China Foreign Language*, 6:57–61.

R. Mitton. 1992. A description of a computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English.

H. T. Ng and Y. S. Chan. 2007. Semeval-2007 task 11: English lexical sample task via english-chinese parallel text. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*.

H. T. Ng and J. K. Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *Proceedings of EMNLP*.

H. T. Ng, B. Wang, and Y. S. Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of ACL*.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

A. Rozovskaya and D. Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proceedings of EMNLP*.

C. C. Shei and H. Pain. 2000. An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13.

M. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of WMT*.

M. Swan and B. Smith. 2001. *Learner English: A Teacher's Guide to Interference and Other Problems*. Cambridge University Press, Cambridge, UK.

J. Tetreault, J. Foster, and M. Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of ACL*.

D. Wible, C. H. Kuo, N. L. Tsao, A. Liu, and H. L. Lin. 2003. Bootstrapping in a language learning environment. *Journal of Computer-Assisted Learning*, 19.

H. Wu and M. Zhou. 2003. Synonymous collocation extraction using translation information. In *Proceedings of ACL*.

J. C. Wu, Y. C. Chang, T. Mitamura, and J. S. Chang. 2010. Automatic collocation suggestion in academic writing. In *Proceedings of the ACL 2010 Conference Short Papers*.

Z. Zhong and H. T. Ng. 2009. Word sense disambiguation for all words without hard labor. In *Proceedings of IJCAI*.