CORRECTING COLLOCATION ERRORS IN LEARNERS' WRITING BASED ON PROBABILITY OF SYNTACTIC LINKS

Azimov A. E. (mitradir@gmail.com)
Lomonosov Moscow State University, Moscow, Russia

Bolshakova E. I. (eibolshakova@gmail.com)
National Research University Higher School of Economics;
Lomonosov Moscow State University, Moscow, Russia

The paper describes a novel method for automatic collocation error correction in NL texts written by language learners or translated from another NL with the aid of machine translators. We assume that the main cause of collocation errors is the strategy of word-by-word translation used by authors of the texts or by machine translators, so the errors essentially depend on the source language. While processing a sentence from the text, the method considers as potential correcting variants all its paraphrases that have the same syntactic structure and are built by replacing all words of the sentence by their substitutes. Substitutes are automatically generated using word translation equivalents taken from a translation dictionary. To detect an error in the sentence, we propose a relevance degree function computed from the probability of the word's syntactic links and applied to the sentence and its paraphrases. If the function value for the sentence is lower than for some of its paraphrases, our method signals an error, then it is corrected by an appropriate sentence paraphrase. The method was evaluated by correcting collocation errors in English texts written by Russian speakers. Stanford Parser and an English text collection were used to gather statistics and compute the probability of English word syntactic links. Within certain limitation, the experiments gave promising results: our method detected about 80% of collocation errors (with words of various POS) and 87% of proposed correcting paraphrases contained a proper correction.

Key words: lexical combinability, collocations, collocation error, error correction, sentence probability, ESL writing

Introduction

Based on computer dictionaries and parsing methods, modern computer text editors and spellers detect both spelling and some syntactic errors in NL texts, but they can't reveal and correct so-called collocation errors. Such errors (e.g., Eng. *heavy tea* instead of *strong tea*; Rus. *играть значение* instead of *иметь значение*) violate norms of lexical combinability. The norms differ for various natural languages and can't be defined by formal rules. As a rule, special collocation dictionaries, for example, [7] include, but only partially, typical collocations for a particular NL.

In modern computational linguistics the concept of collocation may be interpreted in different ways. Following the work [1] we consider a collocation as a stable word combination syntactically related and semantically compatible.

Collocation errors may occur in texts written by native speakers, but more often they arise in texts written by authors for whom the language of the text isn't native, in particular, written by language learners. Such errors also may be the result of automatic translation from one language to another. Below we present several examples of inadequate translation from Russian into English (machine translator Lingvo popular in Russia was used):

- Russian collocation *великий художник* is translated to *great painter*, but in English the collocation *great artist* is more idiomatic;
- Word combination *публичная персона* is translated to *public person* instead of commonly used *public figure*;
- Russian collocation *много денег* is translated into *many money* instead of *much money*.

Such collocation errors often appear in texts as the result of the strategy of word-by-word translation used by either a machine translator or a language learner. The strategy doesn't take into account syntactic and semantic relations between words of the text. The paper [3] argues that the model of native language and the incomplete model of target language (formed in learner's mind or built into the automatic translator) interfere, thus resulting a plenty of mistakes and collocation errors among them.

In the recent decade a number of papers [1, 2, 4, 5, 9, 10] have appeared in the field of computational linguistics, which proposed certain ways to automatically correct collocation errors in NL texts, mainly in English (besides English only Russian and French texts were considered). To detect erroneous collocations the methods proposed in these works use both automatic syntactic analysis of sentences and statistics of word occurrences and cooccurrences. Most of the works employ row statistics, i.e. frequencies of words and word combinations. As a rule, only particular types of word combination are considered while detecting errors: preposition — noun [2], noun — verb [10], collocations with articles [4].

The separate problem is selecting potential word combinations for correcting an erroneous collocation yet detected. Some works, in particular [4], propose to use bases of all possible correcting phrases manually precompiled by human experts, but in practice this way is evidently unreal.

As for accuracy of collocation errors correction, for the works mention above, it varies from 40% [1] up to 69% [10].

This paper describes a method for automatic collocation error correction in NL texts based on probability of word syntactic links computed using statistics of word syntactic links. Unlike previous works, our method handles several types of collocations, including word combinations with content (nouns, adjectives, verbs, etc.) and auxiliary words (prepositions).

While developing the method we suppose the following:

- Texts under correction are written by language learners or are translated from another NL with the aid of certain machine translator.
- The main reason of collocation errors is the applied strategy of word-by-word translation, so the errors essentially depend on the source language.

- Collocation errors don't change the syntactic structure of the sentences; they only substitute erroneous words for correct ones.

Our method follows the idea formulated in [2]: the problem of error correction within a sentence S may be considered as the task to find most probable correcting sentence $\boldsymbol{V}^*$, among possible sentences V, given sentence S:

$$V^* = arg \max_V P(V|S) \tag{1}$$

Taking into account Bayes' theorem, we get:

$$V^* = arg \max_V \frac{P(S|V)P(V)}{P(S)}$$

and then, using conditional independence $P(S)$ of possible sentence V:

$$V^* = arg \max_V P(S|V)P(V) \tag{2}$$

Therefore, to find the most probable substitute sentence V^*, it is necessary to determine probability of sentence V as correcting variant for S and also to determine probability of sentence V.

The text of the paper is organized as follows. First we describe how to build correcting sentences V, we call them *paraphrases*, and how to determine their conditional probabilities $P(S|V)$. Any paraphrase V is constructed by replacing words from S by their *substitute words* that are automatically generated on the bases of word translation equivalents taken from a particular translation dictionary, for example, Russian-English dictionary.

Next, we explain how to determine the probability $P(V)$ of any sentence V in the text (including paraphrases of the source sentence) given its syntactic structure. Further we describe the way how to detect collocation errors in the sentence. For this purpose we propose *a relevance degree function* computed from the probability of word syntactic links and applied to the sentence and their paraphrases. If the function value for the sentence under correction is less than for some its paraphrase, our method signals an error, and the sentence is corrected by appropriate paraphrase.

Then experimental validation of our method is discussed. The method was evaluated by correcting collocation errors in English texts written by Russian speakers. Stanford Parser [6], English text corpora, and Russian-English dictionary were used to compute necessary probabilities. Within certain limitation, the experiments gave promising results.

Finally, we draw conclusions and outline directions for future work.

Paraphrases and their Probabilities

To determine the probability of paraphrases we need to determine the probability of their components, i.e. substitute words.

Recalling our assumption about word-by-word strategy of translation, we define the map *Translate* as set of ordered pairs ⟨*x*,*y*⟩ where word *x* belongs to the source language *X* and has a set of translation equivalents {*y*} from the target language *Y* — cf. Figure 1. We also define the inverse map $Translate^{-1}$ which determines for each word *z* from language *Y* a set of preimages {*x*} from language *X*.

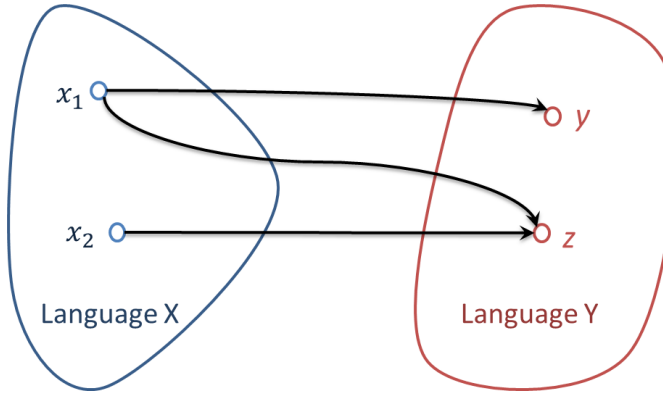

Fig. 1. The map *Translate*

Let $p_+(y|x)$ be conditional probability that word *x* is preimage of word *y* on the map *Translate*. Similarly, let $p_-(y|x)$ be conditional probability on the map $Translate^{-1}$.

In accordance with the assumption that collocation errors arise as a result of word-by-word translation strategy from language *X* to language *Y*, a wrong word , as well as a correct one, both are images of certain word *x* from *X*. For every word *y* we consider *DoubleTranslation* set:

$$DoubleTranslation(y) = \{z \in Y \mid \exists x \in X : z \in Translate(x) \& y \in Translate(x)\} \quad (3)$$

The word *y* is also member of the set.

It is reasonable to consider members of this set as substitute words for every word *y* from language *Y*. However, some members of *DoubleTranslation*(*y*) may have the meaning very different from the meaning of the word *y*. For example, for word *future* (according to Lingvo Russian-English dictionary) the set includes words *next, to be, coming, the beyond, after death, beyond the grave, aftertime, approaching, by-and-by, hereafter, weird*. To overcome this problem we should take into account only synonyms of word *y*, thus obtaining the set:

$$Substitutes(y) = \{z \in Y \mid z \in DoubleTranslation(y) \,\&\, z \in Synonyms(y)\} \quad (4)$$

For example, the set of substitutes for word *beautiful* are *attractive, fine, gorgeous, handsome, pretty.*

Let us consider x as a preimage of words z and y from the language Y, i.e. $y \in Translate\ (x)\ \&\ z \in Translate\ (x)$. The conditional probability of substitute word z, given words y and x, and x is a preimage of y, is defined as follows:

$$p(z|y,x) = p_{-}(x|y)p_{+}(z|x) \tag{5}$$

To compute conditional probability of substitute z for a given word y we must sum values (5) for all common preimages of words z and y:

$$p_{dt}(z|y) = \sum_{\{x|y \in Translate\ (x)\ \&\ z \in Translate\ (x)\}} p_{-}(x|y)p_{+}(x|y) \tag{6}$$

Assuming the independence of collocation errors in the sentence under correction we get the following formula for the conditional probability of paraphrase V, given the sentence S:

$$p(S|V) = \prod_{i} p_{dt}(s_i|v_i),\ s_i \in Substitutues(v_i) \tag{7}$$

Where s_i is a word from the sentence S and v_i is a word of its paraphrase.

Hence we have defined the first factor in formula (2) and need to determine the probability $P(V)$.

Sentence Probability

We build for each sentence its dependency-based parse tree. This tree is a directed graph $G(V, E)$, where V is a set of vertices, they correspond to words of the sentence, and E is a set of edges. Any pair of vertices $(v_1, v_2) \in E$ if and only if the word-vertex v_2 has dependency relation with word-vertex v_1 (v_2 depends on v_1).

The word v_1 is *ancestor* of v_2, if directed path from vertex v_1 to vertex v_2 exists. Let *ancestors*(v_i) denote the set of all *ancestors* for word v_i.

We illustrate the *ancestors* concept with the next sentence: *The main library in the university is one of the largest in Russia*. Its dependency tree is shown in Figure 2, while the Table 1 presents *ancestors* set for each word of the sentence.

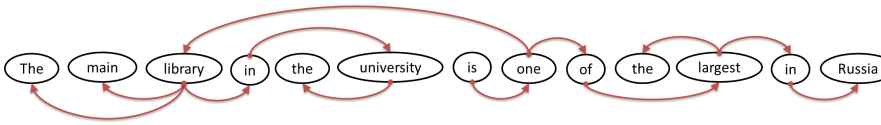

Fig. 2. Example of Dependency Tree

Table 1. Words and corresponding *ancestors* sets

Word	Ancestors
The	library, one, is
main	library, one, is
library	one, is
in	library, one, is
the	university, in, library, one, is
university	in, library, one, is
is	{}
one	Is
of	one, is
largest	of, one, is
the	largest, of, one, is
in	largest, of, one, is
Russia	in, largest, of, one, is

Since we consider collocations as syntactically related word combinations, we assume conditional independence of each word v_i from all other words except its *ancestors*. Thereby the joint probability of the words from the sentence, given the particular sentence parse tree, may be computed as a product of the conditional probabilities:

$$p(v_1, \dots, v_n) = \prod_{i=1}^{n} p\big(v_i | ancestors(v_i)\big) \tag{8}$$

Our method computes probabilities on the bases of word syntactic link statistics gathered on some text collection, so each factor of the above formula is:

$$p\big(v_i | ancestors(v_i)\big) = \frac{N\big(v_i, ancestors(v_i)\big)}{N\big(ancestors(v_i)\big)} \tag{9}$$

where $N(v_i, ancestors(v_i))$ and $N(ancestors(v_i))$ are frequencies of corresponding syntactically related group of words. We should mention that despite of word in the tree root has empty *ancestors* set, the joint sentence probability also depends on root word.

Collocation Errors Correction

Our assumption that collocation errors don't change the syntactic structure of sentences implies that the source sentence *S* and any its paraphrase *V* have isomorphic dependency-base trees — cf. Figure 3. In the source sentence *S* its words s_i were replaced by substitute words v_i (from corresponding *Substitutes* sets), thus resulting paraphrase *V*.

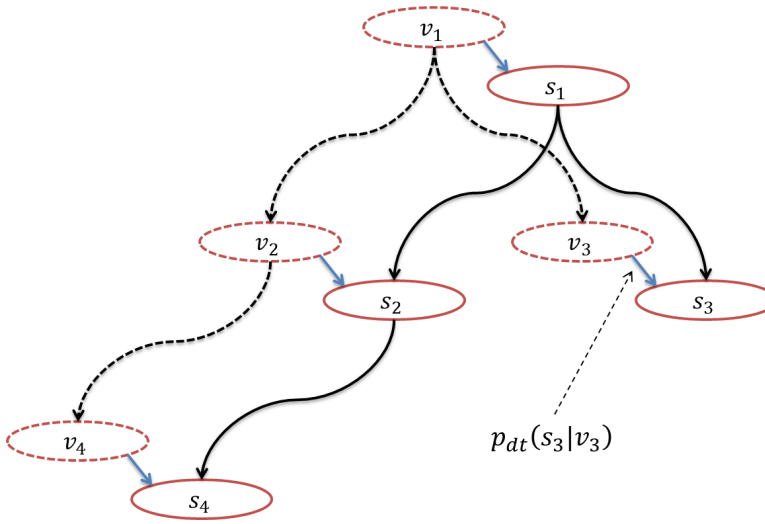

Fig. 3. Isomorphic dependency-base trees

Therefore, the *ancestors* set for word v_i has the following property:

$$ancestors(v_i) = \{v_k \mid v_k \in Substitutes(s_k): s_k \in ancestors(s_i)\}.$$

In such a way, combining formulas (7) and (8), we refine formula (2) and obtain:

$$V^* = arg \max_{v_1,\dots,v_k} \prod_{i=1}^{k} \left(p_{dt}(s_i|v_i)p(v_i|ancestors(v_i))\right) \quad (10)$$
$$\text{where } s_i \in Substitutes(v_i)$$

Next, with the aid of auxiliary coefficient

$$k_{dt}(s_i, v_i) = \frac{p_{dt}(s_i|v_i)}{p_{dt}(s_i|s_i)} \quad (11)$$

For paraphrase evaluation we define a relevance function *Degree*:

$$Degree(v_1, \dots, v_k) = \prod_{i=1}^{k} \left(k_{dt}(s_i|v_i)p(v_i|ancestors(v_i))\right), \quad s_i \in Substitutes(v_i) \quad (12)$$

Using this function, formula (10) could be rewritten as follows:

$$V^* = arg \max_{v_1,\dots,v_k} Degree(v_1, \dots, v_k), \quad s_i \in Substitutes(v_i) \quad (13)$$

Let us give meaningful explanation for Degree function: it measures the correspondence of a particular set of words to a given parse tree. Our method applies the function to the sentence under correction and their paraphrases. If the function value for the sentence is less than for some its paraphrases, an error is detected. In accordance with (13) the detected error is corrected by paraphrase V^* that give maximum Degree value.

We should add that degree of the source sentence is independent of conditional probabilities of substitute words, so the corresponding formula is

$$Degree(s_1, \dots, s_k) = \prod_{i=1}^{k} p(s_i | ancestors(s_i)) \tag{14}$$

Implementation of the Method and Experiments

The described method was evaluated by correcting collocation errors in English texts within the following limitations: there is no more than one collocation error in any sentence of the text, and for each vertex of sentence parse tree only two *ancestors* are considered, namely parent and grandparent.

In this case, the formula (13) can be rewritten as follows:

$$V^* = arg \max_{v_1,\dots,v_k} k_{dt}(s_i|v_i) \prod_{k=1}^{n} p(v_k | ancestors(v_k)) \tag{15}$$

where $v_k = s_k$ if $i \neq k$ and $s_i \in Substitutes(v_i)$ if $i = k$.

The conditional probabilities $p(v_k|ancestors(v_k))$ are based on statistics of syntactic links between words:

$$p(v_i|v_k) = \frac{N(v_i, v_{k1})}{N(v_{k1})} \tag{16}$$

$$p(v_i|v_{k1}v_{k2}) = \frac{N(v_i, v_{k1}, v_{k2})}{N(v_{k1}, v_{k2})} \tag{17}$$

where $N(v_{k1})$ is the frequency of occurrences of word v_{k1};
$N(v_i, v_{k1})$ and $N(v_{k1}, v_{k2})$ are the frequencies of corresponding syntactically linked pairs of words (v_i, v_{k1}) and (v_{k1}, v_{k2});
and $N(v_i, v_{k1}, v_{k2})$ is the frequency of syntactically linked triple of words (v_i, v_{k1}, v_{k2}).

The probability of the parse tree root is determined and calculated as follows:

$$p(v_i) = p(v_i|root) = \frac{N(v_i, root)}{N(root)} \tag{18}$$

In order to conduct experiments, we built a database containing frequencies of syntactically-linked word pairs, triples, and words computed on a large collection of English texts. Stanford Parser [5] was used for automatic parsing of sentences. 220 billion of words were processed; from this data we extracted 18 billion of syntactically linked word pairs and more than 65 billion of syntactically linked word triples. We also used synonyms from Wordnet [8] to compute formula (4).

Since our method is statistical, for error correction we decide to retain, besides the best paraphrase V^* of the source sentence S, all those paraphrases that have the value of *Degree* function greater than the *Degree* value of S. We call the resulted list of paraphrases ordered by *Degree* values *candidate corrections*. They should be suggested for a human editor, in order to make ultimate decision.

The correcting procedure sequentially performs the following steps for each sentence S:

Step 1. Parse sentence S and obtain its dependency parse tree.
Step 2. For each word from S generate its *Substitutes* set and compute conditional probabilities, using the formulas (4) and (6).
Step 3. For each word from S generate a paraphrase V *based* on the generated *Substitutes* set of the word, thus forming a set of paraphrases for S.
Step 4. Calculate the value of function for sentence S by formula (14).
Step 5. Calculate the value of function for all paraphrases V by formula (15).
Step 6. If some paraphrases have *Degree* value that exceeds *Degree* value of S, signal a collocation error.
Step 7. Suggest the ranked list of paraphrases with high values as *candidate corrections* for human editor

For evaluation of our method, we used 70 sentences with typical collocation errors (one error per each sentence) taken from ESL (English as a Second Language) materials. The sentences include erroneous collocations with words of various POS: prepositions, nouns, and adjectives. Besides erroneous sentences ESL materials contain examples of their proper corrections. Some examples of detected erroneous collocations and their proper corrections are presented in Table 2.

Table 2. Detected collocation errors and their corrections

Erroneous sentence	Proper Correction
I think it is a **spend** of my money.	I think it is a **waste** of my money.
To make **understandable**.	To make **plain**.
I have **done** a mistake.	I have **made** a mistake.
The jar was full **with** oil.	The jar was full **of** oil
This is great **painter**.	This is great **artists**.
The **ghost** of the opera.	The **phantom** of the opera.

The experiments showed that 80% of collocation errors were detected. For detected errors, 87% of candidate corrections lists included the proper correction.

In order to analyze the rank of proper corrections within the candidate corrections lists, we choose mean reciprocal rank (*MRR*), which is the arithmetic mean of the inverse ranks of the proper correction in the list:

$$MRR = \frac{1}{L}\sum\frac{1}{r} \tag{19}$$

where L is the number of sentences we used in our experiments, and r is the rank of proper correction in a candidate corrections list. If the list doesn't include proper correction, we assumed that r equals infinity. The Table 3 shows how the number K of corrected sentences depends on the size of candidate correcting list; the last column presents resulting MRR.

Table 3. Results of automatic evaluation MRR

Rank of proper correction	$r = 1$	$r \leq 2$	$r \leq 3$	$r \leq 100$	**MRR**
K	35	45	48	49	0.5

According to presented data the size of candidate correcting list may be equal to 2–3. We should also note that some candidate correcting lists included correct alternative paraphrases that are different from the proper correction.

Conclusions and Future Work

We proposed a novel method for collocation errors correction in learners' writing, based on assumption that the strategy of word-by-word translation is used by authors of the texts. The method automatically generates possible correcting paraphrases, for this purpose it uses translation dictionary from the native language of learners to the language of the text under correction. A relevance degree function was proposed to estimate generated paraphrases and to detect an error; the function is evaluated from the statistics of word syntactic links.

We implemented and evaluated our method supposing only one collocation error in the sentence. The experiments gave promising results: our method detected about 80% of collocation errors and 87% of proposed correcting paraphrases included proper correction.

Directions of our future research are:

- to use Bayesian networks to make the detecting procedure more efficient and test our method on sentences with several collocation errors;
- to expand *Substitutes* sets with word forms and homophones in order to detect additional type of collocation errors.

References

1. *Bolshakova E. I., Bolshakov I. A.* (2007) Automatic detection and computer-aided correction of Russian malapropisms [Avtomaticheskoe obnaruzhenie i avtomatizirovannnoe ispravlenie russkikh malapropizmov], Nauchnaya i Tekhnicheskaya Informatsiya. Ser. 2, No. 5, 2007, p. 8–13.
2. *Brockett C., Dolan W., Gamon M.* (2006) Correcting ESL Errors Using Phrasal SMT Techniques, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics ACL.
3. *Brown P., Pietra S., Mercer R., Pietra V.* (1993) The Mathematics of Statistical Machine Translation, Computational Linguistics, Vol. 19(2).
4. *Gamon M. (2010) Using Mostly Native Data to Correct Errors in Learners' Writing: A Meta-Classifier Approach, Proceeding HLT '10 Human Language Technologies.*
5. *Hermet M., Désilets A., Szpakowicz S.* (2008) Using the Web as a Linguistic Resource to Automatically Correct Lexico-Syntactic Errors, Proceeding The 6th Edition of the Language Resources and Evaluation Conference (LREC 06).
6. *Manning C., Jurafsky D.* Stanford Parser 2012 [html] (http://nlp.stanford.edu/software/lex-parser.shtml).
7. *Oxford Collocations Dictionary* for Students of English. Oxford University Press, 2003.
8. *Wordnet* 2013 [html] (http://wordnet.princeton.edu/)
9. *Wu J., Yu-Chia Chang Y., Teruko Mitamura T., Chang J.* (2010) Automatic Collocation Suggestion in Academic Writing, Proceedings of the ACL 2010 Conference Short Papers.
10. *Yi X., Gao J., Dolan W.* (2008) A Web-based English Proofing System for English as a Second Language Users, Proceedings of IJCNLP.