Developing a Toolkit for Distributional Analysis of Abnormal Collocations in Russian

Polina Panicheva
St. Petersburg State University
St. Petersburg, Russia
ppolin86@gmail.com

Olga Mitrofanova
St. Petersburg State University
St. Petersburg, Russia
oa-mitrofanova@yandex.ru

Abstract

We propose a distributional approach to automatic correction of abnormal collocations in a Russian text corpus containing different types of erroneous word combinations, in particular, construction blending. We develop a toolkit which uses syntactic bigrams from RNC Sketches as training data and Word2Vec semantic model. A corpus of Russian Student Texts with annotation of erroneous word combinations, parsed morpho-syntactically with TreeTagger and MaltParser, was used in experiments. The annotated construction blending errors have been analyzed in terms of error correction by automatically proposing substitution candidates. The correction algorithm involves a set of association metrics based on context selectional preferences and semantic modeling, allowing to rank substitution candidates by their acceptability. Experimental results with nouns annotated as construction blending errors demonstrate the effectiveness of our toolkit. The results show that co-occurrence and Word2Vec semantic models perform ranking of the candidates in terms of different principles: purely constructional and semantic. As a result, the use of Word2Vec semantic filtering improves the quality of error correction.

1 Introduction

The goal of the paper is to model abnormal collocations and correct them automatically. Theoretically abnormal collocation is understood in terms of violation of a syntagmatic relation in a text (i.e., 'You have to try the national ham – jamon'). The abnormal collocation correction model is based on the assumption that a keyword presenting collocation abnormality can be substituted by a word fitting the current context better, while being semantically similar to the initial keyword. In practice, we present an algorithm for automatic correction of abnormal collocations by substituting the keyword with the most frequent word in the given context.

The abmormal collocations are provided by the Corpus of Russian Student Texts (CoRST), (Zevakhina and Dzhakupova, 2015), which consists of educational essays on various topics written by native speakers of Russian. The corpus is annotated, among others, with lexical errors caused by construction blending (Puzhaeva et al., 2015), which involves merging of structural features of different constructions (e.g. ‘*играть роль*’ (*to play the role*) + ‘*занимать место*’ (*to take a seat, to replace*) = *‘*играть место*’ (**to play a seat*)). Blended constructions present a case of abnormal collocations, as they contain at least one word which is untypical in the current context and can be replaced by a semantically similar word to form a proper construction. Moreover, blended constructions present a subtle case of abnormal collocations, as the former are produced by fluent native speakers, and the overall utterance stays meaningful in spite of the blending.

In order to provide a model of abnormal collocation correction we address the following issues:

1. A set of annotated errors by native speakers caused by construction blending is extracted from CoRST;
2. A syntactic-based co-occurrence model is applied to identify and rank substitutes in the blended constructions; a word-embeddings semantic model is added to measure semantic similarity;
3. The construction blending errors are automatically corrected by the proposed model.

2 Related work

Distributional semantic approaches have been applied to identification of a broader scope of lexical anomalies, i.e., metaphor (Shutova, 2010), semantic deviance (Vecchi et al., 2015) and learner errors (Kochmar and Briscoe, 2013). We follow (Shutova, 2010) in applying the Context-Based Paraphrasing weighting algorithm to identifying and ranking possible substitution candidates. However, the difference of our work is that CBP is based on collocation counts with individual words for error identification, while Shutova (2010) analyzes word clusters for a more abstract metaphoric usage. Kochmar and Briscoe (2013) apply features derived from word-embeddings semantic models to identify learner errors, whereas in our case word-embeddings have to be combined with a fine-grained syntax-based CBP model to handle subtle errors.

A common feature of the mentioned works is that training and test data are constructed from relatively frequent keywords of English, and context words are added according to a restricted list of syntactic relations (attribute, verb subject or/and object). The crucial difference of the current work is that both training and test data consist of unrestricted corpora containing all possible syntactic relations, thus rendering the task closer to a real-life problem. This causes obvious difficulties, such as word and collocation sparsity and imbalance issues. There are also important restrictions imposed by Russian NLP resources, with the syntactic bigram statistics available only in terms of SynTagRus syntactic relations (Boguslavsky et al., 2002), restricting the relevant morpho-syntactic algorithms to TreeTagger and MaltParser (Sharoff et al., 2008; Sharov and Nivre, 2011). It is also noteworthy that current training and test corpora belong to different genres, rendering the task genre-independent.

Our work is the first attempt to automatically approach an unrestricted (by frequency or syntactic properties) corpus of real-world lexical errors. To our knowledge, it is the first approach to lexical anomalies in Russian texts by native speakers. The datasets of syntactic and distributional variety have been applied for model training. The novelty of the method involves combining syntactic count-based and word-embeddings distributional models.

3 Toolkit design

The aim of the toolkit is to analyze, correct and, to an extent, identify word collocations, with anomalies in contextual restrictions caused by creative language processing in metaphor, violation of fine-grained selectional restrictions in the texts of language learners and native speakers, pronounced mistakes caused by speech impairment. The basic assumption is that a coherent text complies with the requirements concerning semantic and selectional restrictions on syntagmatic relations between words. Technically it is rendered by the idea that a word basically occurs in contexts in which it has already occurred frequently, or in some sense similar ones. The system thus learns co-occurrence regularities from text corpora and processes a keyword and its context, measuring their mutual association and proposing substitutes for the keyword where possible.

The toolkit is expected to work in two settings. First, it should provide analysis and substitutes for words annotated as abnormal in a text. This setting is applied in the current work. Second, it should be able to automatically identify some abnormal words in collocations. The latter goal is a subject of future work.

3.1 Input

Fine-grained selectional restrictions analysis requires either very large datasets or syntactic processing. In Russian, morphological analysis is required in both settings. Bag-of-Words models, as Word2Vec, do not require any further parsing, but offer paradigmatic-oriented insight which is difficult to interpret and tune in a syntagmatic collocational setting. While syntactically parsed corpora are difficult to obtain, they provide fine-grained information which is indispensable when identifying the nature of syntagmatic violation.

As a training corpus we use the RNC Sketches syntactic bigram statistics[1]. It provides statistics on syntactic relations based on a sample of the Russian National Corpus (RNC) of 200M words, where every keyword is associated with a list of its relations and their frequencies. A syntactic relation is a pair *(relation, word)*, where the relations inventory is that of the SynTagRus corpus (Boguslavsky et al., 2002), and the word is the dependent word, e.g. '*попробовать (try) -> 1st completive -> себя (oneself) : 126*', '*попробовать (try) -> 1st completive -> блюдо (dish) : 7*', '*национальный (national) -> attrib -> идея (idea) : 390*', '*национальный (national) -> attrib -> блюдо (dish) : 55*'. An additional step of reverting the syntactic relations is required to obtain source words for every dependent keyword. In order to unify the format of the training data and the data used for error analysis,

[1] http://ling.go.mail.ru/synt/

we apply MaltParser and TreeTagger used to create RNC Sketches (Sharoff et al., 2008; Sharov and Nivre, 2011) to the testing data. Individual word frequencies were obtained from the Russian Frequency Dictionary (Lyashevskaya and Sharov, 2009). We also supply our algorithm with a Word2Vec semantic model based on RNC (Kutuzov and Andreev, 2015).

The data used for automatic error analysis is provided by the Corpus of Russian Student Texts (CoRST). It contains educational texts by native speakers of Russian (500K words) annotated with a broad range of errors (10K annotated errors). The errors caused by construction blending (Puzhaeva et al., 2015) are especially relevant to our task, as they present subtle violations of selectional restrictions.

3.2 Statistical models

We use the RNC Sketches syntactic bigrams as a syntactic model and apply automatic ranking of the erroneous keywords based on their context. The list of possible substitutes for a particular keyword is generated as the list of words occurring in the bigram corpus in the same syntactic context as keyword. Namely, it is the intersection of the words occurring with every syntactic relation in the keyword context. The substitutes are commonly ranked using the following association measure scores: Mutual Information scoring (Khokhlova, 2008), context-based paraphrasing (CBP) (Shutova, 2010), Resnik's selectional association based on Kullback-Leibler distance (Resnik, 1993), and Word2Vec-based semantic scoring (Kutuzov and Andreev 2015). The likelihood *L* of a particular paraphrase *i* of the word *w* is estimated as the likelihood of the joint events: the substitute *i* co-occurring with all the other lexical items from its context $w_1, \dots w_N$ in syntactic relations $r_1, \dots r_N$.

Context-Based Paraphrasing: The context-based paraphrasing likelihood estimation is based on syntactic co-occurrence:

$$L_i(CBP) = \frac{\prod_{n=1}^{N} f(w_n, r_n, i)}{(f(i))^{N-1}}.$$

Word2vec Semantic Scoring: In order to account for purely semantic word properties, i.e. restrict the list of substitutes to words semantically similar to the keyword, we apply the Word2Vec model trained with RNC data. Semantic similarity between a keyword *kw* and it's substitute *i* is calculated as the cosine distance between the corresponding vectors in the Word2Vec semantic space:

$$Sim(kw, i) = \cos(kw, i).$$

4 Experimental setup

We perform a proof-of-concept experiment by automatically correcting the errors caused by construction blending in CoRST with context-based paraphrasing and additional Word2Vec semantic scoring. The errors are made by native speakers and represent violations of selectional restrictions. There are 130 lexical errors in the corpus caused by construction blending. We have extracted 29 sentences from the corpus, containing a noun annotated as a lexical error caused by construction blending. We set out to automatically suggest a list of substitutes for the erroneous nouns and score them according to the Context-Based Paraphrasing procedure. We also perform Word2Vec semantic filtering to improve the results.

№	Example sentence	Syntactic context		Weighted substitutes		Evaluation result	
		Relation	Word	Candidate	Likelihood	Strict	Loose
1	Между нравами и законами трудно провести четкое **различие**. - It's hard to draw a strict **difference** between customs and laws.	1st completive	провести - draw	**линия - line** **грань - border** **разграничение - distinction** **граница - boundary**	82.5 60.4 49.1 42.9	Corr	Corr
		attrib	четкий - strict				
2	Обязательно попробуйте национальный **окорок** – хамон, … - You have to try the national **ham** – jamon, …	1st completive	попробовать - try	сила - power **блюдо - dish** напиток - drink **продукт - product**	21.0 12.3 9.5 2.7	Inc	Corr
		attrib	национальный - national				
3	приходится платить за каждый аттракцион и из-за их дорогой **стоимости**… - one has to pay for every attraction, and because of their high **price** …	prepos	из-за - because of	черта - feature отношение - relation страх - fear лес - forest	0.02 0.0009 0.0008 0.0004	Inc	Inc
		quasi-agent	они - they				
		attrib	дорогой- high				

Table 1. Examples of context-based paraphrasing results.

We calculate the accuracy of the results by applying manual evaluation. A substitute candidate is marked correct if it fits the context better than the erroneous keyword and leaves the meaning of the sentence unchanged. The resulting lists of candidates contain up to 50 ranked words. The assumption is that the highest ranked words represent the best substitution candidates in the provided contexts. It is examined by manually analyzing a short-list of top candidates. Evaluation is performed in two settings:

1. The **strict mode** implies that the substitutes provided by the algorithm are correct if the candidate with the highest rank is correct.
2. The **loose mode** renders the substitutes list correct if there is a correct candidate among the four highest ranked candidates.

We do not perform further evaluation procedures at this stage, because the initial proof-of-concept experiment is aimed at providing an overall insight on the task, its restrictions and improvement possibilities.

5 Results and discussion

5.1 Context-based paraphrasing

Out of 29 sentences, 4 contained morphological and syntactic annotation errors in the morpho-syntactic analysis of the erroneous nouns, which made the list of the candidates provided by CBP empty. The rest of the examples, 25 sentences, were processed with CBP substitute ranking.

Out of 25 examples, the algorithm provided **15 (60%)** correct substitutes in the loose mode and **10 (40%)** in the strict mode. The results of the substitution experiment are exemplified in Table 1. Analysis shows that among the 10 loose-mode incorrect results, 5 are defined by the syntactic context which doesn't allow retrieving any meaningful candidates: there is a very limited number of candidates co-occurring with all the context features in the corpus, and their meaning is either too broad or too distant from that of the original keywords (for example, in '*это было обусловлено православной религией*' (*it was preconditioned by orthodox religion*) substitutes for '*религия*' (*religion*) only include '*образование*' (*education*), '*организация*' (*organization*)). However, strict mode-specific mistakes include correct substitutes, which are downgraded in their rank by the words fitting the syntactic context very well but bearing a meaning unrelated to the keywords (see ex. 2 in Table 1). These cases could be improved by adding purely semantic information to the model.

5.2 Semantic filtering

Shutova (2010) performs semantic filtering based on WordNet by limiting the paraphrasing candidates to those in hypernym or co-hyponym relations with the keyword restricted to three-level distance. In order to avoid sparsity of data covered by hand-coded resources, we apply RNC-based Word2Vec model as a semantic filter to eliminate substitution candidates unrelated to the keyword. The semantic similarity threshold

№	Example sentence	Weighted substitutes	
		No filtering	Word2Vec semantic filtering
1	… Между нравами и законами трудно провести четкое **различие**. - It's hard to draw a strict **difference** between customs and laws.	**линия - line** **грань - border** **разграничение - distinction** **граница - boundary**	**грань - border** **разграничение - distinction** **граница - boundary** параллель - parallel
2	Если рассматривать этот вопрос с религиозной **стороны** то тут тоже тяжело найти оправдание. – Looking at the issue from the religious **side**, …	**точка - point** **позиция - position** начало - start язык - language	**точка - point** **позиция - position** конец - end
3	Поэтому отдых на Байкале … помогает человеку снова набраться жизненной **силой**. – Holiday at Baikal … helps one to collect life **power**.	опыт - experience впечатление - impression **энергия - energy** дух - spirit	опыт - experience **энергия - energy** дух - spirit мудрость - wisdom
4	… круглые сироты, не имеющие в целом **свете** ни единого родственника? – … total orphans, having no relatives in the whole **world(1)**?	ряд - row **мир – world(2)** арсенал - arsenal район - region	**мир – world(2)** **жизнь - life** страна - country город - city
5	Обязательно попробуйте национальный **окорок** – хамон, … - You have to try the national **ham** – jamon, …	сила - power **блюдо - dish** напиток - drink **продукт - product**	**блюдо - dish** напиток - drink **продукт - product** **лакомство - delicacy**
6	… люди, ставящие перед собой высокие **рамки** – … people who set high **limits**	**цель - goal** оценка - mark честь - honour точка - point	**цель - goal** **планка - bar** барьер - barrier положение - position

Table 2. Differences between the results with and without semantic filtering.

value is experimentally set to **0.1**.

As expected, filtering results in slight improvement in loose mode evaluation, correctly analyzing **18** examples (**72%**). However, it gives considerably higher results in the strict mode, eliminating semantically unrelated candidates and ranking correct substitutes higher: accuracy evaluated in the strict mode is **14** examples **(56%)**. Table 2 illustrates the meaningful differences between the substitution results with and without semantic filtering, with the keywords in the example sentences and the correct substitutes highlighted in bold and the results crucial for the strict mode performance also underlined. It is important to notice that semantic filtering also improves the performance beyond the first-rank candidate by making qualitative modifications to the candidate list: it reduces the number of low-likelihood substitutes (ex. 2), increases the rank of correct substitutes and their proportion in the four highest-ranked candidates (ex. 3, 5, 7).

6 Conclusions and future work

We have introduced a toolkit for abnormal collocation analysis and automatic correction. The toolkit applies collocation-based association measures aimed at analyzing various types of context restriction violations. We have performed a proof-of-concept experiment with construction blending errors by native speakers of Russian, which confirms the applicability of the statistical association measures to this task. Close analysis of algorithm errors has revealed the need for semantic restrictions, which cannot be accounted for by purely context-based methods. Adding Word2Vec-based semantic filtering has improved the results qualitatively and in terms of accuracy, making the incorporation of various language models a promising approach in analyzing abnormal associations. Another crucial point in this task is accurate and consistent morpho-syntactic analysis of training and test corpora.

Our future work includes adding more data to the analysis (other parts of speech annotated in CoRST) and processing anomalies of a different nature: learner errors, intentional semantic deviance in figurative language, errors caused by language impairment.

7 Future considerations

An important finding of the current experiment is the need for combination of fine-grained syntactic and distributional semantic models. The combination is expected to play a crucial role in future analysis of different error types. As shown in current research, native speaker errors present subtle co-occurrence violations while basically maintaining the meaning of the keyword comparing to its correct substitute. However, we expect a different trade-off between syntactic co-occurrence and semantics in other types of errors. It appears that the higher the level of a language learner, the more the erroneous combinations maintain their basic meaning; whereas the lack of immediate experience with fluent text is reflected in co-occurrence violations, regardless of language proficiency level.

Figurative text has been shown to contain semantic violations of a specific type, as in metaphor, where the meaning of a source domain is projected onto a different target domain (Shutova, 2010). Metaphor presents errors violating the basic semantic restrictions, but requiring a more abstract semantic analysis based on word clusters and domains. On the contrary, speech impairment is expected to produce semantic violations with no underlying abstract pattern or with a pattern fundamentally different from that identified in figurative language.

Acknowledgments

The reported study is supported by RFBR grant 16-06-00529 "Development of a linguistic toolkit for semantic analysis of Russian text corpora by statistical techniques".

References

Igor Boguslavsky, Ivan Chardin, Svetlana Grigorjeva, Nikolai Grigoriev, Leonid Iomdin, Leonid Kreydlin, and Nadezhda Frid. 2002. Development of a dependency treebank for Russian and its possible applications in NLP. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, vol. III, pages 852–856

Maria Khokhlova. 2008. Extracting collocations in Russian: Statistics vs. dictionary. In *JADT 2008: 9es Journ´ees internationales dAnalyse statistique des Donn´ees*, pages 613–624.

Ekaterina Kochmar and Ted Briscoe. 2013. Capturing anomalies in the choice of content words in compositional distributional semantic space. In *RANLP*, pages 365–372.

Andrey Kutuzov and Igor Andreev. 2015. Texts in, meaning out: neural language models in semantic similarity task for Russian. *arXiv preprint arXiv:1504.08183*.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL (2)*, pages 302–308.

Olga Lyashevskaya and Sergey Sharov. 2009. Chastotnyy slovar'sovremennogo russkogo yazyka (na materialakh natsional'nogo korpusa russkogo yazyka) [The frequency dictionary of modern Russian (on the materials of the Russian National Corpus)]. *Moscow: Azbukovnik Publ*.

Svetlana Puzhaeva, Natalia Zevakhina, and Svetlana Dzhakupova. 2015. Construction blending in non-standard variants of Russian in the Corpus of Russian Student Texts. In *Proceedings of the 6th International Conference "Corpus Linguistics-2015"*, 390-397. Saint-Petersburg. (in Russian)

Philip Stuart Resnik. 1993. Selection and information: a class-based approach to lexical relationships. *IRCS Technical Reports Series*, page 200.

Serge Sharoff, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a Russian tagset. In *LREC*.

Sergey Sharov and Joakim Nivre. 2011. The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge. In *Proceedings of the Annual International Conference Dialogue, Computational Linguistics and Intellectual Technologies*, number 10, page 657.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037. Association for Computational Linguistics.

Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2015. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces.

Natalia Zevakhina and Svetlana Dzhakupova. 2015. Corpus of Russian student texts: design and prospects. In *Proceedings of the 21st International Conference on Computational Linguistics "Dialog"*. Moscow, 2015.