Distributional analysis of Russian lexical errors

Polina Panicheva
Department of Mathematical Linguistics, Saint Petersburg State University, Russia
https://doi.org/10.36505/ExLing-2016/07/0029/000288

Abstract
An algorithm of analyzing obscure lexical collocations is proposed. It is based on a co-occurrence model and distributional semantic filtering. We apply the proposed technique to lexical errors of construction blending, as annotated in the Corpus of Russian Student Texts. Results of error processing are analyzed and classified; reasons for different results in the paraphrasing experiment are discussed.
Keywords: Distributional Semantics, lexical errors, construction blending, Russian.

Introduction
We propose a framework for analyzing violation of syntagmatic relations resulting in construction blending [Puzhaeva et al. 2015]. Our toolkit includes models of meaning and selectional restrictions, applied to analyzing different types of abnormal collocations: native speakers' and learners' errors, metaphorical expressions, peculiarities in clinical texts, etc. The algorithm allows to identify and correct obscure collocations. We discuss the application of our approach to a corpus of native speaker errors.

Datasets
As a training corpus we use the RNC-Sketches syntactic bigram statistics. It provides statistics on syntactic relations in the Russian National Corpus (RNC), where every keyword is associated with a list of its relations and their frequencies in terms of MaltParser and TreeTagger; the latter are used to create RNC Sketches [Sharoff 2008, Sharov 2011] to the testing data. Total word frequencies were obtained from the Russian Frequency Dictionary [Lyashevskaya, Sharov 2009]. We supply our algorithm with an RNC-based Word2Vec semantic model [Kutuzov, 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 and is annotated with different types of 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.

Statistical models

We use the RNC-Sketches syntactic bigrams as the syntactic model and apply automatic ranking of the erroneous keywords based on their context. The list of possible substitutes for a particular keyword is the intersection of the words occurring with every syntactic relation in the keyword context. The substitutes are ranked using the association measure scores: context-based paraphrasing (CBP) [Shutova 2010], and Word2Vec-based semantic scoring [Kutuzov, Andreev 2015].

Context-based paraphrasing

The context-based paraphrasing (CBP) likelihood estimation is based on the same grounds of syntactic co-occurrence, but is not symmetric and does not account for context word frequencies:

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

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:

$$(5)\ Sim(kw, i) = \cos(kw, i)$$

The similarity threshold for the candidates with the initial erroneous word is experimentally set to 0.1.

Experiment setting

We perform a proof-of-concept experiment by analyzing 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 27 sentences in the corpus annotated with a noun presenting a lexical construction blending error. We set out to automatically suggest a list of substitutes for the erroneous nouns and score them according to the CBP procedure with Word2Vec semantic filtering.

The results are manually analyzed, and the errors are grouped according to their proposed substitution candidates. The first group contains errors for which the distributional algorithm proposed no

relevant candidates. For the second group we calculate the Accuracy of the results by applying manual evaluation. A candidate is marked correct if it fits the context at least as well as the erroneous keyword and leaves the meaning of the sentence unchanged. 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.

Results and analysis

Errors with no substitution candidates

There are 12 errors with no relevant candidates proposed. Eight of them obtain candidates by CBP, but the candidates are correctly filtered out by Word2Vec. Four errors get no proposed candidates, as their syntactic context is so obscure that there are no words attested in the corpus occurring with all the relevant syntactic distribution. The errors are exemplified in Table 1. Manual analysis shows that all of these cases appear to contain no error, or the error is annotated with a mistake, e.g. for a wrong word (Ex.1). A few of the 11 cases contain morphosyntactic analysis errors (Ex.1) or obscure syntactic relation names (Ex.2) immediately affecting the CBP candidate choice.

Errors with relevant substitution candidates

15 errors obtain substitution candidates with CBP which pass the semantic filtering. Examples are presented in Table 2. Nine errors are correctly analyzed in the strict mode (Ex.1), 4 errors are correctly analyzed in the loose mode (Ex.2). There are 2 errors left which only get incorrect candidates (Ex.3).

Conclusions

The distributional approach to lexical errors is an adequate measure of the distributional specificity of a construction in text; it also presents a useful tool which automatically suggests lexical substitutes for unusual lexical co-occurrences. Where lexical substitution is impossible, manual analysis confirms no lexical error in the sentence (44%). Proposed lexical substitutes (56%) are correct in 60% and 87% in strict and loose mode respectively.

Future work includes modifying the morphosyntactic analysis to minimize parsing errors. Future applications of the approach include

specific error collections, i.e. language acquisition and learner errors, clinical texts, in order to shed light on their distributional nature.

Table 1. Errors with empty candidate lists.

№	Example sentence	Syntactic context	
		Relation	Word
1	…всех тех, кто взял на себя **роль** донесения фактов до массового сознания / … those who took the **role** of informing the masses	1-компл	взять / take
		1-компл	донесение / informing
		до_Gen	сознание / -
2	находит **себе** применение третий ход по реализации стратегии дискредитации… / the third approach to discredit applies **itself** …	неакт-компл	находить / -

Table 2. Errors with relevant substitution candidates.

№	Example sentence	Candidates	Result
1	Обязательно попробуйте национальный **окорок** – хамон... / You have to try the national **ham** – jamon…	блюдо / meal напиток / drink продукт / product	Strict correct
2	Если следовать **взглядам** Мари Биша … / following the **views** of Marie Bichat …	тенденция / trend правило / rule	Loose correct
3	Путешествие в Санкт-Петербург не нанесет **ущерба** вашему кошельк / A trip to St. Petersburg will not bring **damage** to your purse	потеря / loss	Incorrect

Acknowledgements

The reported study is supported by RFBR grant 16-06-00529.

References

Kutuzov, A., Andreev, I. 2015, 'Texts in, meaning out: neural language models in semantic similarity task for Russian', arXiv preprint arXiv:1504.08183.

Lyashevskaya, O., Sharov, S. 2009, The Frequency Dictionary of Modern Russian (on the materials of the Russian National Corpus), Moscow. (in Russian)

Puzhaeva, S.; Zevakhina, N., Dzhakupova, S. 2015, Construction blending in non-standard variants of Russian in the Corpus of Russian Student Texts. Proc. 6th Intern. Conf. "Corpus Linguistics-2015", 390-397. St. Petersburg. (in Russian)

Sharoff, S.; Kopotev, M.; Erjavec, T.; Feldman, A., Divjak, D. 2008, Designing and Evaluating a Russian Tagset., in 'LREC'.

Sharov, S., Nivre, J. 2011, The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge. Proc. Annual Intern. Conf. Dialogue, Computational Linguistics & Intellectual Technologies', 657.

Shutova, E. 2010, Automatic metaphor interpretation as a paraphrasing task, in 'Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL', pp. 1029--1037.