Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams

Kormacheva, Daria
University of Tübingen
2014


http://hdl.handle.net/10138/153191

Downloaded from Helda, University of Helsinki institutional repository.
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
Abstract

This paper focuses on automatic determination of the distributional preferences of words in Russian. We present the comparison of six different measures for collocation extraction, part of which are widely known, while others are less prominent or new. For these metrics we evaluate the semantic stability of automatically obtained bigrams beginning with single-token prepositions. Manual annotation of the first 100 bigrams and comparison with the dictionary of multi-word expressions are used as evaluation measures. Finally, in order to present error analysis, two prepositions are investigated in some details.

1 Introduction

In this paper we present our ongoing research on the distributional preferences of words and their co-occurrences in Russian.

Our research follows the tradition of distributional analysis, which takes its roots in the work of Harris (1951). The core idea of this approach is that the semantic similarity/dissimilarity between words correlates with the distributional properties of their context. The most known line of this research is distributional semantics, which is based on the assumption that “at least certain aspects of the meaning of lexical expressions depend on the distributional properties of such expressions, i.e. on the linguistic contexts in which they are observed” (Lenci, 2008). In theory, the distributional properties should be studied on all language levels, including phonetics, prosody, morphology and syntax, semantics, discourse, and pragmatics (Gries, 2010). In practice, however, some properties are more difficult to obtain than others; as a consequence, researchers usually focus on a limited amount of linguistic phenomena.

In particular, multi-word expressions (MWEs), in which a given word participates, form the immediate context of this word; the distributional properties of such context can be used for word categorization and description. However, this immediate context is not homogeneous; it is formed by MWEs of various semantic nature: idioms, multi-word lexemes, collocations, i.e. “co-occurrences of words”, and colligations, i.e. “co-occurrence of word forms with grammatical phenomena” (Gries and Divjak, 2009).

Distinguishing all these types of MWEs is not a simple task, since there is no clear boundary between them. For example, a word combination can be simultaneously a collocation and a colligation – in (Stefanowitsch and Gries, 2003) this type of MWE is called collostruction. Goldberg (2006) proposed that language as such is a constructicon, with fusion being its core nature. Thus, measuring the strength of grammatical and/or lexical relations between words is not a trivial task.

The situation becomes even more complicated for morphologically rich languages, because each word may have several morphological categories that are not independent and interact with each other.

In our project we aim to implement the model able to process MWEs of various nature on an equal basis. It compares the strength of various possible relations between the tokens in a given n-gram and searches for the “underlying cause” that binds the words together: whether it is their morphological categories, or lexical compatibility, or both.

Our research is motivated by the recent studies on grammatical profiling, including those by Gries and Divjak (2009), Gries (2010), Janda and Lyashevska (2011), Divjak and Arppe (2013).
These works are focused on classification of the certain classes of words using profiles, i.e. distributions of grammatical and lexical features of the context. A profile does not necessarily include all the context features, but only those for which the word has some distributional preferences. This selectivity, as Janda and Lyashevskaya (2011) fairly point out, is the crucial part of the methodology since “it is necessary to target precisely the level of granularity at which the interaction between linguistic category and morphology (or other formal structure) is most concentrated”.

The main difference between these works and our study is that these researchers establish the proper level of granularity before the main phase of the analysis, while one of our main goals is to extract these profiles from the corpus. As has been mentioned before, we try to implement a unified model; the set of input queries for such a model is unrestricted and, as a consequence, the profiles cannot be set a priori.

For example, Janda and Lyashevskaya (2011) have shown that tense, aspect and mood form a sub-paradigm for Russian verbs, while person, number and gender are not relevant to this interaction. However, they have found that, for instance, a particular class of verbs – rude ones – has a significant preference of the singular number in imperfective imperative form. This demonstrates that no language property can be excluded from analysis beforehand.

In the previous stage of this project, (Kopotev et al., 2013), we mainly dealt with colligations. We have developed an algorithm that takes as an input an n-gram, in which one position is an unknown variable, and finds the most stable morphological categories of the words that can fill this gap. An in-depth evaluation focusing on a limited number of linguistic phenomena, namely bigrams beginning with single-token prepositions, has been conducted.

In this paper we continue to investigate the same material, i.e. Russian bigrams that match the [PREPOSITION + NOUN] pattern. Our particular task is to analyse MWEs, which are extracted with the help of our algorithm and can be free or stable to various extents. The n-gram corpus, extracted from a deeply annotated and carefully disambiguated sub-corpus of the Russian National Corpus is used as the data. The size of this corpus is 5 944 188 words of running text.

## 2 Method

In general, our system takes any n-gram of length 2-4 with one unknown variable as an input and tries to detect the most stable word categories that can stay for this variable. These categories include token, lemma and all morphological categories of the Russian language. The initial query pattern may contain various constraints, for example, number or tense can be specified for the unknown variable. Alternatively, the pattern can be unrestricted and formed only by the combination of the surrounding words.

The most stable lexical and grammatical features for a given query pattern are defined using normalized Kullback-Leibler divergence. The category with the highest value of normalized divergence is considered to be the most significant for the pattern. The detailed algorithm and evaluation of the first results can be found in (Kopotev et al., 2013).

Obviously, the most stable grammatical feature for the [PREPOSITION + NOUN] pattern is case that has maximal divergence for all prepositions. The next step is to determine the exact values of the category; i.e., continuing this example, the particular cases that can co-occur with the preposition. Note, that due to unavoidable noise in corpus annotation, the model cannot return all the values found in the data.

Dealing with grammar, we use simple frequency ratio that is able to find possible cases for each preposition with reasonably high quality: precision 95%, recall 89%, $F_1$-measure 92% (Kopotev et al., 2013). However, frequency ratio does not demonstrate such a performance on detecting stable lexical units.

In this paper we use various statistical measures to extract collocations from raw text data and analyse the obtained results. The following measures we applied:

**frequency:** $f(p, w)$, where $p$ is the pattern, $w$ is the wordform that can appear within this pattern, $f(p, w)$ is the absolute frequency of the wordform in the pattern.

**refined frequency ratio:**

$$FR(p, w) = \frac{f(p, w)}{f(w)}$$

, where $f(w)$ is the absolute frequency of the wordform in the general corpus. The grammatical categories of the wordform are taken into account, because its surface form can be ambiguous.
For example, many Russian nouns have the same form in nominative and accusative cases; if such a word occurs within the pattern in accusative case, we use only accusative case to count its corpus frequency.

**weighted frequency ratio**

which is a frequency ratio multiplied by logarithm of the word frequency in the general corpus:

\[ wFR(p, w) = FR(p, w) \times \log f(w) \]

The idea behind this measure is as follows. Let us consider two words, \( w_1 \) that appears in the corpus 2 times and \( w_2 \) that appears in the corpus 1000 times. Let \( f(p, w_1) = 1, f(p, w_2) = 500 \); hence, \( FR(p, w_1) = FR(p, w_2) = 0.5 \). It is obvious that the \( w_1 \) may appear within the pattern by accident, whereas the fact that \( w_2 \) occurs within the pattern 500 times out of 1000 is meaningful. We multiply the frequency ratio by logarithm of the word frequency to give more weight to frequent words.

Finally, we compare these three measures with the following widely used metrics:

**mutual information**, \( MI \)  
(Church and Hanks, 1990):

\[ MI(p, w) = \log \frac{f(p, w)}{f(p) \times f(w)} \]

**Dice score**,  
(Daudaravicius, 2010):

\[ dice(p, w) = \frac{2 \times f(p, w)}{f(p) + f(w)} \]

**t-score**,  
(Church et al., 1991):

\[ t-score(p, w) = \frac{f(p, w) - f(w) \times f(p)}{\sqrt{f(p) \times f(w)}} \]

Thus, 6 different measures are used for the evaluation in this paper: part of them are widely known, while others are less prominent or new.

3 Experiments and Results

We evaluate the semantic stability of automatically obtained bigrams beginning with single-token prepositions. We investigate 25 prepositions, such as “без” (without), “в” (in/to), etc. For each preposition, algorithm collects all the bigrams that match the pattern [PREPOSITION + \( w \)], where \( w \) is a noun. In order to minimize noise in our data, bigrams containing infrequent nouns with \( f(w) > 5 \) are filtered out.

The remaining bigrams are sorted according to the aforementioned statistical measures, which means that for each preposition 6 different rankings are presented. We then compare these rankings to determine the most appropriate statistical measure. Such a comparison becomes itself a tricky task since no “gold standard”, i.e. no complete list of collocations, is available. In this paper we perform two types of evaluation: comparison with the dictionary of multi-word expressions (Rogozhnikova, 2003), and manual annotation of the first 100 bigrams in each ranking.

3.1 Comparison with the dictionary

**Explanatory dictionary of expressions equivalent to word** (Rogozhnikova, 2003) contains approximately 1500 Russian MWEs. These expressions have various nature and can behave as either lexical or function words. They are not necessary idiomatic in terms of semantics, and their only common property is stability: they have the constant form that allows little or no variation.

In particular, the dictionary contains a vast amount of expressions with prepositions, including complex adverbs, prepositions and conjunctions, as well as idiomatic expressions. They constitute the most comprehensive list of Russian MWEs with prepositions, which is crucial for our current task.

For each ranking, we calculate the **uninterpolated average precision** (Moirón and Tiedemann, 2006; Manning and Schütze, 1999): at each point \( c \) of the ranking \( r \) where a dictionary entry \( S_c \) is found, the precision \( P(S_1...S_c) \) is computed and all precision points are then averaged:

\[ UAP(r) = \frac{\sum_{S_c} P(S_1...S_c)}{|S_c|} \]

The uninterpolated average precision (UAP) allows us to compare rankings and indirectly measures recall (Manning and Schütze, 1999). Results, showing the UAP for each ranking, are presented in Table 1; we report the results for 17 prepositions only, because the dictionary does not contains any entries for the rest.

It can be seen from the Table 1 that simple frequency is the most appropriate measure to determine fixed expressions and idioms; other frequency-based measures, namely weighted frequency ratio and t-score, demonstrate comparable performance, while the refined frequency ratio, Dice-score and MI are not appropriate for this task.

The possible explanation may be the fact that the dictionary contains many MWEs, equivalent to prepositions, conjunctions or adverbs. It has been shown before, (Yagunova and Pivovarova, 2010), that MI is more appropriate to extract topical units.
of the corpus – such as complex nominations, terminology and noun groups that are significant for a particular document – while t-score tends to extract pragmatic units, which characterize the corpus in general.

3.2 Manual Annotation

The dictionary-based evaluation, presented in the previous section, cannot be considered a complete one. Although the high ranks of dictionary MWEs probably mean that for these expressions the ranking should be considered relevant, we cannot tell anything certain about other bigrams in the ranking. One obvious reason is that for many prepositions there are no entries in the dictionary. For example, although every native speaker of Russian knows the idiom “кроме шуток” (joking apart) (literally “all jokes aside”), the dictionary contains no fixed expressions for nouns with the preposition “кроме” (beyond/except). Moreover, as it is always the case with the dictionaries, some fixed expressions can be neglected in the list.

Furthermore, fixed expressions and idioms are not the only object of our study. Many MWEs do not fulfil the aforementioned requirement of stability; distributional preferences, which our model should be able to catch, do not necessary lead to the lexical rigidity of the expression.

Thus, in this section we present the second evaluation, based on the manual annotation of the extracted bigrams. The first 100 bigrams in each ranking were manually annotated and each bigram was categorized either as a fixed expression/idiom or as a free word combination. Then the interpolated average precision was calculated (see the formulae presented in the Section 3.1). Results, presenting the UAP for each ranking, are shown in the Table 2.

It can be seen from the table that the results we got for the manual annotation are quite similar to those obtained for the dictionary-based evaluation. As before, Dice score and MI proved to be not suitable for this task; frequency-based measures, namely frequency, weighted frequency ratio and t-score again demonstrated approximately the same performance. The refined frequency ratio performed slightly worse than these three measures, although in general the number of collocations obtained using this measure is higher than for the dictionary-based evaluation.

On the whole, these results can be considered negative: the first 100 bigrams extracted using the best statistical measure – weighted frequency ratio – in average contain less than 25% of fixed expressions and idioms. But despite the average low performance of the algorithm, it is worthy to note that there is a high variety among the prepositions. For the results based on manual evaluation and sorted according to the weighted frequency ratio, the UAP varies between 0 and 73.34. This can be partially accounted for by the fact that various Russian prepositions have different tendency to form fixed expressions. Below we will illustrate this on the example of two prepositions.

4 Error Analysis and Discussion

In order to perform error analysis, we investigate the following prepositions: “без” (without) and “в” (near/at). These prepositions were selected since for “без” (without) our method achieved the best result (73% of the bigrams extracted using wFR contain fixed expressions and idioms), while “в” (near/at) was among the prepositions, for which our method failed. Nevertheless, these two prepositions have a common feature that can be used to improve the performance of our algorithm in the future.

The bigrams restrained by both prepositions are often part of various constructions. Among the first 100 nouns extracted by wFR for preposition “без” (without), 11 are parts of the construction “[без]+piece of clothing]: “галстук” (tie), “перчатка” (glove), “пого” (epaulette), “шапка” (cap), etc.; 3 are included into construction related to the formalities at border checking points: “виза” (visa), “паспорт” (passport), “штамп” (stamp).

The same holds for the first 100 nouns extracted by wFR for preposition “в” (near/at). Nouns obtained for this pattern may be described in terms of the following constructions:

10: “[в]+relative: “ребенок” (child), “папа” (dad), “гепха” (mother in law), etc.;
16 bigrams: “[в]+part of house: “окно” (window), “крыльцо” (porch), “стена” (wall), etc.;
6: “[в]+nationality: “немец” (German), “русский” (Russian), “цыган” (Gypsy), etc.

We may see that such constructions constitute
Preposition | Meaning | \( f \) | \( rFR \) | \( wFR \) | \( MI \) | \( dice \) | \( t \)
--- | --- | --- | --- | --- | --- | ---
без | without | 33.16 | 33.89 | 35.58 | 1.45 | 1.14 | 30.60
в | in/into | 24.94 | 14.64 | 29.55 | 0.59 | 2.33 | 24.90
dля | for | 3.12 | 0.17 | 0.42 | 0.37 | 0.07 | 4.41
до | until | 26.95 | 27.74 | 38.67 | 0.85 | 0.71 | 25.44
за | behind | 22.62 | 25.56 | 53.13 | 0.17 | 0.16 | 23.06
из | from | 1.28 | 0.86 | 1.43 | 0.18 | 0.10 | 1.27
из-за | from behind | 33.33 | 29.17 | 50.00 | 0.42 | 0.37 | 33.33
к | to | 34.62 | 3.19 | 24.84 | 0.25 | 0.23 | 34.75
между | between | 25.00 | 0.27 | 0.56 | 0.38 | 0.16 | 25.00
на | on | 12.58 | 8.32 | 7.83 | 0.72 | 0.47 | 11.85
от | from | 16.01 | 1.98 | 5.15 | 0.25 | 0.16 | 15.60
перед | in front of | 50.00 | 0.35 | 0.98 | 0.37 | 0.18 | 50.00
по | by/up to | 35.83 | 16.72 | 34.36 | 1.44 | 0.98 | 35.22
под | under | 31.50 | 20.04 | 21.73 | 0.10 | 0.86 | 31.13
при | at/by | 43.99 | 8.77 | 43.08 | 0.75 | 0.34 | 43.99
про | about | 25.00 | 7.69 | 20.00 | 0.18 | 0.18 | 20.00
с | with | 13.20 | 7.63 | 16.85 | 0.59 | 0.58 | 13.22

Average: 25.48 | 12.18 | 22.60 | 0.59 | 0.53 | 24.93

Table 1: The number of fixed expressions from the dictionary among Russian \([\text{PREPOSITION} + \text{NOUN}]\) bigrams. For each preposition we present the uninterpolated average precision for all bigrams sorted according to the following measures: \( f \) – frequency, \( rFR \) – refined frequency ratio, \( wFR \) – weighted frequency ratio, \( MI \) – mutual information, \( dice \) – Dice score, \( t \) – t-score.

| Preposition | Meaning | \( f \) | \( rFR \) | \( wFR \) | \( MI \) | \( dice \) | \( t \)
--- | --- | --- | --- | --- | --- | --- | ---
беz | without | 72.86 | 68.38 | 73.34 | 7.17 | 5.83 | 72.60
в | in/into | 47.93 | 35.14 | 58.40 | 7.87 | 4.33 | 49.37
для | for | 7.28 | 12.32 | 14.69 | 0.13 | 0.42 | 7.26
до | until | 44.03 | 52.38 | 60.93 | 0.00 | 0.00 | 44.37
за | behind | 38.58 | 44.90 | 51.58 | 13.11 | 5.36 | 38.7
из | from | 4.48 | 7.84 | 12.29 | 0.00 | 0.00 | 4.63
из-за | from behind | 10.06 | 10.90 | 11.47 | 0.00 | 0.00 | 9.97
из-под | from under | 6.60 | 12.37 | 8.92 | 6.72 | 8.19 | 5.99
к | to | 11.99 | 0.97 | 22.28 | 2.43 | 3.19 | 23.49
кроме | beyond/except | 5.18 | 3.68 | 5.18 | 0.00 | 0.00 | 5.18
меж-ду | between | 9.28 | 5.18 | 5.18 | 2.00 | 1.88 | 9.25
на | on | 23.95 | 39.52 | 25.32 | 10.10 | 10.16 | 34.
на-д | above | 0.48 | 0.00 | 0.00 | 0.00 | 0.00 | 0.48
о | about | 0.98 | 0.00 | 0.00 | 0.00 | 0.00 | 0.87
от | from | 15.81 | 10.71 | 11.06 | 0.00 | 0.00 | 15.94
перед | in front of | 11.69 | 0.00 | 0.33 | 0.00 | 0.00 | 11.69
по | by/up to | 57.50 | 43.29 | 60.18 | 7.89 | 7.89 | 57.35
под | under | 62.68 | 57.95 | 62.69 | 0.15 | 0.01 | 62.29
при | at/by | 32.49 | 11.30 | 19.49 | 9.65 | 7.01 | 32.49
про | about | 3.08 | 2.06 | 3.08 | 0.00 | 0.00 | 3.08
ради | for | 24.32 | 20.11 | 22.14 | 3.58 | 4.03 | 23.22
с | with | 36.34 | 30.97 | 44.11 | 0.57 | 0.75 | 36.69
у | near/at | 3.97 | 1.92 | 4.17 | 0.00 | 0.00 | 2.92
через | through | 4.93 | 3.23 | 5.06 | 8.59 | 5.82 | 4.94

Average: 22.35 | 19.80 | 24.25 | 3.33 | 2.70 | 23.24

Table 2: The number of fixed expressions among Russian \([\text{PREPOSITION} + \text{NOUN}]\) bigrams. For each preposition we present the uninterpolated average precision for the first 100 bigrams sorted according to the following measures: \( f \) – frequency, \( rFR \) – refined frequency ratio, \( wFR \) – weighted frequency ratio, \( MI \) – mutual information, \( dice \) – Dice score, \( t \) – t-score.
a considerable part of the extracted bigrams. Just to illustrate the point, counting these bigrams as relevant collocations would increase the UAP for “без” (without) from 73.34% to 85.47% and for “у” (near/at) from 4.17% to 73.82%. Similar observations can be done for other prepositions in our list.

Thus, we must re-think the initial problem statement and aim to not only extract fixed expressions and idioms for a given query pattern, but also deal with the kind of expressions described above. We should further define what is the status of such MWEs as at the counter, at the TV-set, at the window, etc. These are not fixed expressions in a sense. Their meaning can be inferred from the meanings of the parts, and the pattern is productive. Nevertheless, these expressions still have something in common and can be described in terms of constructions that predict some grammatical and semantic features of a word class. So we can suppose that in this case the choice of the collocate is not accidental either. This assumption returns us back to the initial point of this article. The model would be a more accurate representation of natural language, if it deals with collocations rather than with two separate classes of collocations and colligations.

Practically, we assume that such constructional preferences can be found by similar algorithms if the corpus is semantically annotated. If we would have semantic annotation at our disposal, we would be able to group words according to their semantic tags (e.g., animal, relative or nationality) and extract different kinds of constructions in the same way as we do with other categories. Unfortunately, our data do not contain any semantic annotation and we do not have access to any Russian corpus suitable for this task. Still, in our future work, we will try to bootstrap semantic classes from the data on the grounds of the same procedure of distributional analysis.

Acknowledgements

We are very grateful to M. Pierce and R. Yangarber for their support and contribution to the preparation of this paper. Also we would like to thank E. Rakhilina, O. Lyashevskaia and S. Sharoff for providing us with the data for this research.

References


