

ВЛИЯНИЕ СИНТАКСИЧЕСКОЙ СТРУКТУРЫ НА ИЗВЛЕЧЕНИЕ КОЛЛОКАЦИЙ-СУЩЕСТВИТЕЛЬНЫХ ПРИ ГЛАГОЛАХ

Акинина Ю. С. (jakinina@hse.ru),
Кузнецов И. О. (iokuznetsov@hse.ru)

Центр семантических технологий НИУ ВШЭ, Москва, Россия

Толдова С. Ю. (toldova@yandex.ru)

МГУ имени М. В. Ломоносова; Центр семантических технологий НИУ ВШЭ, Москва, Россия

Ключевые слова: коллокации, глагольная сочетаемость, автоматический синтаксический анализ, корпусные методы

THE IMPACT OF SYNTACTIC STRUCTURE ON VERB-NOUN COLLOCATION EXTRACTION

Akinina Y. S. (jakinina@hse.ru),
Kuznetsov I. O. (iokuznetsov@hse.ru)

The Center for Semantic Technologies, Higher School of Economics, Moscow, Russia

Toldova S. Y. (toldova@yandex.ru)

Lomonosov Moscow State University, The Center for Semantic Technologies, Higher School of Economics, Moscow, Russia

Automatic verb-noun collocation extraction is an important natural language processing task. The results obtained in this area of research can be used in a variety of applications including language modeling, thesaurus building, semantic role labeling, and machine translation. Our paper describes an experiment aimed at comparing the verb-noun collocation lists extracted from a large corpus using a raw word order-based and a syntax-based approach. The hypothesis was that the latter method would result in less noisy and more exhaustive collocation sets. The experiment has shown that the collocation sets obtained using the two methods have a surprisingly low degree of correspondence. Moreover, the collocate lists extracted by means of the window-based method are often more complete than the ones obtained by means of the syntax-based algorithm, despite

its ability to filter out adjacent collocates and reach the distant ones. In order to interpret these differences, we provide a qualitative analysis of some common mismatch cases.

Keywords: collocations, verb compatibility, parsing, corpus methods

1. Introduction

The identification of semantically related words is an actual NLP task. Most of the special lexicographic research in that area is focused on identifying synonymy, multi-word expressions or hyperonymy-hyponymy relations. One special case of lexical relation extraction is modeling verb semantics using the information about the verb-noun compatibility. The information obtained from the verb-noun distribution model is then used in a wide range of NLP tasks such as semantic-role labeling ([Gildea, Jurafsky 2002]), word sense disambiguation ([Kustova, Toldova 2009], fact extraction, thesaurus building ([Lin 98], [Pado, Lapata 2007]), machine translation ([Orliac, Dillinger 2003]) and others. Modeling lexical relations between words can be done automatically by methods of collocation extraction.

Generally, the collocation extraction is a two-step process which includes candidate extraction and candidate ranking. A variety of methods are proposed for executing each of these steps. In particular, the candidate selection can be performed using either linear or syntactic text representation. In the first case, the words in the source texts are treated as consequent units, while the second model relies on syntactic representation in which units are connected non-linearly. Being more complex from a technical point of view, the latter method should result in noise reduction and higher recall due to the additional syntactic filtering. Our aim is to compare the collocation sets obtained by using these two candidate extraction methods in the same setting, and to analyze the differences between the results.

2. Background

2.1. Notion of collocation

The definition of “collocation” differs across linguistic traditions. From the theoretical point of view, collocation can be considered to be some kind of a “fixed phrase”, in a scale where fixed phrases are opposed to “free phrases” (see [Khokhlova 2008] for a review). However, when it gets to practice, retrieving a particular theoretically predetermined class of phrases can become problematic. A more practical definition of collocation within the corpus linguistics paradigm will be “two or more lexical units that co-occur more often than would be expected by chance” [Manning, Schütze 1999]. Statistically-based methods of collocation extraction ranks word pairs according to a certain measure of association, which evaluates

the chance of their occurrence. As a consequence a high rank can be obtained not only by fixed expressions such as *сломать голову* “rack one’s brains”, but by free word combinations such as *сломать руку* “to break smb’s hand”. While the former is an idiom listed in a dictionary the meaning of the latter is compositional. However the noun *рука* in the latter is a “typical” argument for the verb *сломать*. This type of word-combinations should be also taken into consideration for verb semantics modeling. Thus we use the term collocation for both types of word combinations discussed above.

2.2. Collocation candidate selection methods

Choosing the method of compiling lists of possible collocates is a crucial step in collocation extraction. The variety of methods can be roughly divided in two groups. The methods from the first group, which we refer to as **window-based methods** (e. g. [Church, Hanks 1990], [Breidt 1993], [Todirascu et al. 2008], [Todirascu, Gledhill 2008]), rely on linear word order model, in which the collocation candidates are extracted from a fixed-size window, and the distances correspond to the raw distance between two (or more) words as presented in the source document. Analyzing adjacent bigrams can be regarded as a particular case with window size of 1. Applying POS or pattern filters can be implemented as an approximation to syntactic structures ([Klyshinskij et al. 2010], [Todirascu et al. 2008]). The second group can be referred to as **syntax-based methods** ([Lin 1998], [Kilgariff, Tugwell 2002], [Khokhlova 2009]). The methods from this group rely on syntactic structure instead of using the linear representation. The candidate list is generated based on syntactic relations. Both candidate selection methods have advantages and shortcomings. Window-based methods tend to extract additional noisy data and ignore the long-distance syntactical links ([Kilgariff, Tugwell 2002]), but are easy to implement. Using the syntax-based methods makes it possible to filter out spurious examples in the nearest context and also access the distant collocates, which are invisible in the window-based linear representation ([Kilgariff, Tugwell 2002]). The increase in precision comes at the cost of carefully describing all the syntactical constructions, in which two collocation candidates can occur.

2.3. Verb-noun collocation extraction

When it comes to verb-noun (V-N) collocations, the researchers’ aim is usually to extract some specific, theoretically predetermined types of constructions ([Breidt 1993], [Todirascu et al. 2008], [Todirascu, Gledhill 2008]). There is a particular interest among researchers for the task of V-N collocation extraction. The majority of the works are based on the combination of morphological pattern-based and statistically based methods (see [Todirascu et al. 2008] for French and Romanian, [Breidt 1993] for German, [Todirascu, Gledhill 2008] for English and Romanian, [Todirascu et al. 2008] for German). For this method a high level of noise is reported (c. f. [Todirascu, Gledhill 2008]). On the one hand, a certain amount of totally irrelevant V-N pairs were extracted by means of the method.

On the other hand, the authors were looking for some particular types of collocations. For instance, subject+predicate collocations or combinations with circumstantial adjunct ([Todirascu et al. 2008], [Todirascu, Gledhill 2008]) were out of the author's interest. Therefore, the conclusion was that the syntactic information was required to detect such combinations ([Todirascu et al. 2008]). The authors of [Breidt 1993] also claim that syntactic parsing is necessary to distinguish subject-verb from object-verb combinations. Indeed, applying syntactic filters over a parsed corpus in English allows getting statistical information from Subject-Verb-Object triples to accurately answer the questions about typical arguments, e.g. "*What can you drink?*" [Church, Hanks 1990].

There are some works that focus on collocation extraction in Russian, but so far the main method has consisted of applying/comparing different word association measures over the lists of adjacent word units (e.g. [Khokhlova 2008], [Jagunova, Pivovarova 2010]). The experience of using syntax-based Word Sketches methodology presented in [Khokhlova 2009] claims viability of this method for collocation search in Russian, but does not analyze Verb-Noun collocations in particular. The work presented in [Klyshinskij et al. 2010] concerns extracting verb lexical compatibility information in general (that is, not just fixed phrases, but typical free phrases as well), but it relies on a huge text corpus to bypass the syntactic parsing using a number of assumptions (such as "the next noun phrase after a single verb most probably depends on it"). The authors of [Klyshinskij et al. 2010] report a good correlation between the high-frequency part of the list and the lists obtained with collocation methods.

To sum up, the majority of works on Verb-Noun collocations investigate the nouns that are involved in some particular types of verb-argument relations or Verb-Noun fixed Expressions (excluding Klyshinskij et al. 2010). In our research all the syntactically related to a verb noun are taken into consideration irrespective of their syntactic role (direct object, circumstantial NP etc.).

3. Experiment

3.1. Setup

The goal of the experiment described below is to compare the verb-noun collocations extracted from a large corpus with and without use of syntactic information. The corpus was preprocessed with a tokenizer, a POS-tagger, a morphology analyzer and a syntactic parser. We use the **Pointwise mutual information (PMI)** as a statistic measure for verb-noun collocation extraction. The two methods for collocation candidates extraction are used: the first one is a window-based bag of words method, the second one is based on the results of syntactic parsing.

The resulting collocation sets were grouped by verb and compared in order to evaluate the degree of correspondence between the extracted sets. The results of this comparison were then analyzed in detail, and several conclusions about the mismatching cases were made.

3.2. Corpus

In order to obtain sufficient amounts of initial data, a roughly 9 million word corpus of Russian newspaper texts was used¹. The corpus consists of planarized random sentences sampled from various news articles published in the period from April 2011 to April 2012. This results in a certain lexical skewness, as the vocabulary, describing the events which have taken place in that period of time, influences the word distribution over the corpus. However, we believe that this factor can be disregarded, taking into account that the corpus was used to compare two automatic methods on the same dataset without use of any external data.

3.3. Preprocessing

The corpus has been preprocessed using a set of tools developed by S. Sharoff and J. Nivre ([Sharoff, Nivre 2011]), which includes a tokenizer, a TreeTagger-based ([Schmid 1994]) part-of-speech tagger, a lemmatizer based on CSTLemma ([Jongejan, Dalianis 2009]), and a Russian dependency parser model for the MaltParser ([Nivre et al. 2006]) trained on SynTagRus syntactic corpus ([Boguslavsky et al. 2000]). The preprocessed texts have been allocated in a relational database and indexed. The resulting dataset consists of tokens which are mapped to words; each word is assigned a set of morphological features and a lemma, and for each sentence a set of labeled dependency relations is fixed. S. Sharoff and J. Nivre report 95–97% POS-tagger accuracy and an unlabeled attachment score of 88 ([Sharoff, Nivre 2011]), which seems sufficient for our type of analysis, taking into account that we aim to extract statistically dependent word pairs from a large corpus, and the impact of accidental errors should be smoothed by the dataset size. Although the relation labels are available, they weren't used during the experiment, so all the dependency links were treated as unlabeled ones.

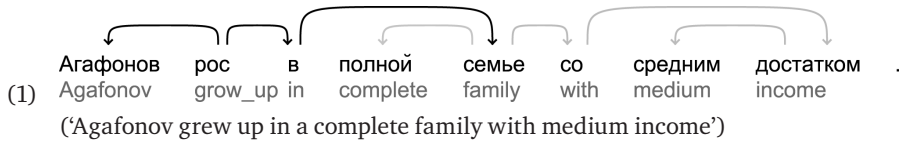
3.4. Collocation extraction

We have extracted collocations from the syntactically parsed corpus using two different strategies for obtaining the initial collocation candidate lists. Only finite verb forms were analyzed due to the fact, that the non-finite forms in Russian often lack some of the overtly expressed arguments, so taking these forms into account would require additional transformation and preprocessing steps.

The first candidate extraction strategy is to build potential collocate pairs by extracting unlabeled verb-noun dependency relations. In case of prepositional objects where the dependency relation points at a preposition, the preposition was skipped (see the collocation candidates *расту* ('grow up') — *Агафонов* ('Agafonov'))

¹ The corpus was collected by H. Christensen and is available on <http://corpora.heliohost.org/>

и *расти* ('grow up') — *семья* ('family') in the example (1)). The total of 358,915 verb-noun pairs was obtained. We will refer to the collocations resulting from these pairs as syntax-based or dependency-based collocations.



The second strategy is to use a window-based approach. In order to estimate the appropriate window size, the distribution of distances between verbs and their dependent nouns was analyzed (see **Fig. 1**) and the window size of $[-5; 5]$ words was selected as a result.

For every finite verb form in the corpus, we have extracted all the nouns found in the same sentence in the context of $[-5; 5]$ words. The non-word tokens such as punctuation and numbers were ignored. From all the extracted pairings, a collocation candidate list containing verb lemma, noun lemma and the collocation frequency was formed. 708,131 collocations were extracted using the bag-of-words strategy. We refer to the collocations obtained using this method as **window-based collocations**.

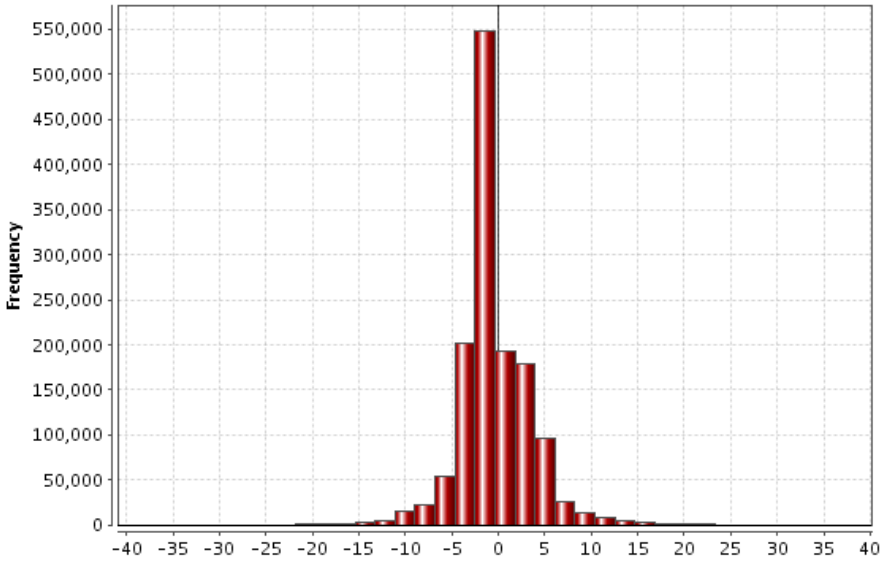


Fig. 1. Verb-argument distance distribution

The collocation candidates were ranked using the PMI metric, which is defined as

$$pmi(x, y) = \log \frac{P(x, y)}{P(x) * P(y)}$$

PMI as a word association measure has several drawbacks, among them an overrating of infrequent combinations and a poor accordance with expert collocation lists evaluation ([Evert, Krenn 2001]). The first is usually handled by establishing frequency cutoff thresholds. In our experiment only the combinations containing verbs with total raw frequency more than 100 and nouns with total raw frequency more than 10 were analyzed. As for the combination frequency threshold itself, we have found out that the cutoff value of 10 filters out too many combinations, resulting in very short final sets as compared to the sets obtained using lower cutoff threshold. Lower cutoff thresholds introduce some noisy data but also increase the recall, e. g.:

(2) *сломать* ('break')

c10wc10 syntax, window: рука ('arm'), нога ('leg')

c5wc5 syntax, window: нога ('leg'), нос ('nose'), ребро ('rib'), рука ('arm')

c2wc2 syntax: нога ('leg'), нос ('nose'), ребро ('rib'), результатом ('result'),
рука ('arm'), челюсть ('jaw')

c2wc2 window: андрей ('Andrej'), бедро ('hip'), год ('year'),
женщина ('woman'), камера ('camera'), лицо ('face'), мальчик ('boy'),
матч ('match'), нога ('leg'), нос ('nose'), падение ('fall'),
палец ('finger'), побои ('beating'), раз('once'), ребро ('rib'),
результат ('result'), рука ('arm'), челюсть ('jaw'), шея ('neck')

The Cn notation is used to denote the cutoff threshold of n for syntactic model while the WCn denotes the cutoff threshold n for window-based model. For example, c10wc5 means that thresholds of 10 and 5 were applied to syntax- and window-based models respectively.

We have varied the frequency cutoff thresholds to examine the changes in correspondence between collocate sets built by using dependency-based and window-based approach. We have also compared the lists obtained using unequal thresholds.

3.5. Evaluation

The candidate sets extracted for each verb were ranked by PMI, and only the top 20 collocates were selected. In order to evaluate the degree of correspondence between the lists obtained using syntax- and window-based methods, for each verb we have calculated two weighted intersection measures using the formula:

$$WI(A; B) = \frac{|x \in (A \cap B)|}{|x \in A|}$$

Let window be the set of collocations for a given verb extracted using the window-based technique. Let syntax be the one extracted using the syntax-based method. The measure aims to describe how good the window-based list of nouns fits into the one extracted using the syntax-based representation. The measure is inversely the ratio of words from the syntax-based list, which are also presented in the list of words obtained by applying the window method. These measures can be thought of as Precision and Recall with syntax-based set treated as Key. As with Precision and Recall, the harmonic mean of two measures (F-measure) was also computed using the standard formula:

$$F_1 = \frac{2 * WI(window, syntax) * WI(syntax, window)}{WI(window, syntax) + WI(syntax, window)}$$

The comparison results are presented in Table 1.

Table 1. Comparison of window-based and syntax-based collocations

WI(window,syntax)			WI(syntax>window)				
	wc10	wc5	wc2		wc10	wc5	wc2
c10	0,62109	0,27844	0,11761	c10	0,93730	0,84038	0,60494
c5	0,79878	0,55038	0,20042	c5	0,60583	0,88075	0,64143
c2	0,67867	0,66600	0,49630	c2	0,22842	0,44996	0,69669
average=0,48974			average=0,65396				

F1-measure			
	wc10	wc5	wc2
c10	0,71847	0,38428	0,17429
c5	0,66367	0,64752	0,26928
c2	0,30926	0,51122	0,55542

3.6. Results

The evaluation shows a moderate level of correspondence between the results obtained by comparing the two methods discussed. As WI measures for different combinations of minimal combination frequency threshold shows (Table 1) using distant threshold values (e.g. c10-wc2) leads to the worst results. According to the table, the best F_1 is achieved using the threshold of 10 for both syntax- and window-based algorithms, though the small size of resulting sets must be taken into account. The best $WI(syntax>window)$ is achieved by using cutoff threshold of 5 on syntax- and the one of 10 on window-based candidates.

The value of $WI(\text{syntax}, \text{window})$ averaged on all threshold combinations is significantly higher than the one of $WI(\text{syntax}, \text{window})$. This reflects the fact that in many cases the majority of the words from syntax-based lists are included into the window-based lists, while the opposite is false.

Taking into account the starting hypothesis and presuming that the syntactic relations should give less noisy data, one could suggest that the collocation sets, obtained using window-based candidate list, would lack precision and introduce too much noisy data. However, the expert analysis shows that in many cases the collocates extracted by the window-based model are perfectly relevant to the task and should be treated as correct. Consider the following examples from sets obtained using frequency threshold of 5 in both algorithms. Matching words are shown in bold, and relevant words are underlined.

(3) **забить** ('kick, score') c5wc5

syntax: **ворота** ('goal'), **год** ('year'), **гол** ('goal'), **голова** ('head'),
матч ('match'), минута ('minute'), мяч ('ball'), сезон ('season'),
тревога ('alarm'), форвард ('forward'), шайба ('puck')

window: **ворота** ('goal'), **год** ('year'), **гол** ('goal'), **голова** ('head'),
игра ('game'), команда ('team'), матч ('match'), минута ('minute'),
момент ('moment'), мяч ('ball'), пенальти ('penalty'),
полузащитник ('halfback'), сезон ('season'), смерть ('death'),
состав ('members'), счет ('score'), тайм ('time'), тревога ('alarm'),
чемпионат ('championship'), шайба ('puck')

4. Discussion

The analysis of the results shows that both methods share some common advantages and disadvantages, and the particular disadvantages of each method can be both due to experimental setting drawbacks and linguistic features of the texts. It turns out that, contrary to the expectations, the window-based method tends to extract some relevant verb-noun collocations, which are absent in the sets obtained by the syntax-based method. While the window-based approach also results in a higher level of noise, the syntax-based method suffers from narrowness of syntactic patterns used to extract collocation candidates. Our results show that using simple syntactic patterns is insufficient to model the semantic relations between predicate verbs and their arguments, which results in lower recall.

4.1. Common shortcomings

4.1.1. Corpus skewness

A common shortcoming of the lists obtained using both candidate extraction techniques is collocation specificity, which is related to the skewness of the source data. The texts in our corpus were obtained from news articles released in a one-year

period, so the names of objects which were often mentioned in the media in that time span influence the statistics obtained from the whole corpus. That problem could be partially solved by using a larger and more representative corpus or recognizing and filtering out named entities.

(4) возглавить ('be head of') c10wc10

syntax: год ('year'), рейтинг ('rating'), совет ('council'), список ('list')

window: александр ('Alexander'), владимир ('Vladimir'), год ('year'), группа ('group'), дмитрий ('Dmitry'), комитет ('committee'), медведев ('Medvedev'), отделение ('department'), партия ('party'), правительство ('government'), президент ('president'), путин ('Putin'), рейтинг ('rating'), руководитель ('leader'), сергей ('Sergej'), совет ('council'), список ('list'), управление ('board'), человек ('man')

4.1.2. Capturing the parts of other constructions

In some cases, the verb is syntactically related to a head of a fixed expression. In this case the collocations extracted by both methods will be invalid (see the examples below):

(5) следить ('follow') c5wc5

syntax: ход ('progress')

window: ход ('progress'), голосование ('voting')

(6) 
 Как член комиссии, следил за ходом голосования на дому.
 As member comission monitor process voting at home
 ('As a member of the commission, he followed the progress of the voting at home')

4.2. Window method disadvantages

4.2.1. Capturing the dependant of a valid collocate

These cases are similar to the example (7), but here the collocation extracted by the syntactic method may be considered valid. At the same time, the window-based approach erroneously extracts its dependant:

(7) отклонить ('decline'):

syntax: жалоба ('complaint'), иск ('suit'), предложение ('proposition'), суд ('court')

window: жалоба ('complaint'), иск ('suit'), москва ('Moscow'), предложение ('proposition'), суд ('court')

(8) 
 Арбитражный суд Москвы вчера отклонил иск Росимущества к ЗАО.
 Arbitrary Court Moscow yesterday decline suit Rosimushestvo against JSC
 ('Yesterday the Moscow Arbitrage Court declined the suit of Rosimushestvo to JSC...')

The collocations extracted this way reflect the skewness of the corpus.

4.2.2. Frequent uninformative noise

The occurrence of high-frequency non-informative words like *человек* ('man'), *год* ('year'), *Россия* ('Russia') is a prominent feature of the collocation lists extracted by window-based method. They may be linguistically unrelated, as *человек* in (9):

(9) **отпустить** ('let out')с5wc5

syntax: залог ('bail'), игрок ('player'), свобода ('freedom'), суд ('court')

window: залог ('bail'), игрок ('player'), свобода ('freedom'), суд ('court'), человек ('man')

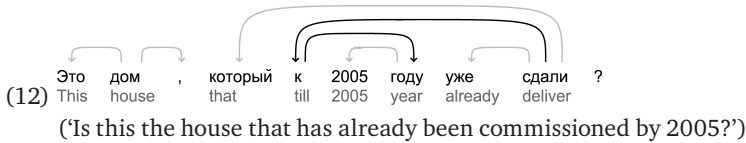


It may also be a member of a regular circumstantial construction as *год* in (11):

(11) **сдать** ('pass')с5wc5

syntax: экзамен ('exam')

window: год ('year'), экзамен ('exam')



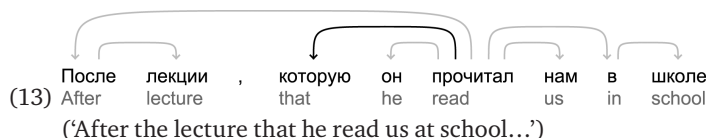
This type of nouns in the top of the window lists is due to the general frequency of expressions of an event time in a clause (it's also true for some other semantic relations). For instance, the collocation *год* ('year') is found only by window-based method in 137 verbs out of 548 within the cutoff threshold of 5. We suppose that an additional procedure of filtering such cases could increase the degree of syntax-based and window-based lists overlapping.

4.3. Syntax-based method shortcomings

Although we have analyzed only finite verb forms in order to reduce syntactic complexity, there are still many issues related to describing the syntactic construction in which semantic relatedness can be expressed. In many cases, a related noun was not captured by the syntax-based candidate extraction algorithm due to the absence of the direct syntactic relationship to the verb in a sentence. Common cases include relative clauses, argument coordination and object pronominalization.

4.3.1. Relative clauses

One common case which is not taken into account by our syntax-based model is the one when the verb is located in a relative clause as in the following example:



The window-based model was able to extract the candidate *лекции* (*lectures*) + *прочитать* (*read*) while the power of our syntax pattern-based method was insufficient to capture the semantic relatedness between these two words. In cases when the amount of such constructions is high, this issue can influence the overall corpus statistics, e.g.:

- (14) прочитать ('read') c10wc10
syntax: интернет ('Internet'), книга ('book')
window: интернет ('Internet'), книга ('book'), лекция ('lecture')

4.3.2. Argument coordination

Another syntactic relation type that should be taken into account is coordination. The parser that we used is based on the framework where the dependency relation between a verb and its coordinated arguments is drawn to the first of these arguments, followed by a chain dependency through a conjunction. See Figure 16, where the verb "выехали" (go, leave) is connected only to the first argument *полицейские* (*policemen*). That first argument is then connected to the second one, *сотрудники* (*officials*) with the conjunction *и* (*and*).

- (15) выехать ('drive off') c5wc5
syntax: автомобиль ('car'), группа ('group'), место ('place'), полоса ('lane'), раз ('once')
window: автомобиль ('car'), глава ('head'), год ('year'), группа ('group'), движение ('traffic'), дом ('house'), машина ('car'), место ('place'), область ('region'), полиция ('police'), полоса ('lane'), происшествие ('accident'), сотрудник ('official'), управление ('board'), человек ('man')



Note that the window-based method succeeds to extract collocations in some of these cases.

4.3.3. Argument pronominalization

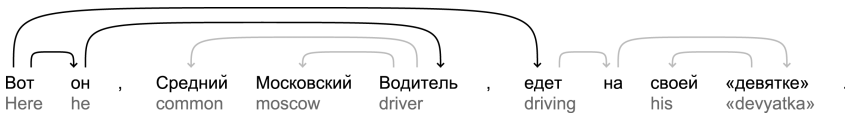
The final drawback of the syntactic method which is worth mentioning is that our model lacks co-reference information. In many cases, the core arguments of a verb

(especially, the subject and object) are substituted by a pronoun. When the antecedent is in the same sentence, it still can be located by the window-based approach, but the syntax-based candidate extractor fails to identify the candidate due to the lack of coreference information, as in the following example:

(17) *ехать* ('go, travel')c5wc5

syntax: вагон ('carriage'), машина ('car')

window: автобус ('bus'), вагон ('carriage'), водитель ('driver'), год ('year'), машина ('car'), минута ('minute'), человек ('man')

(18) 

 Вот он , Средний Московский Водитель , едет на своей «девятке» .
 Here he common moscow driver driving на his «devyatka»
 ('Here he is, the Average Moscow Driver, traveling in his "devyatka" (car model)')

However, the antecedent is not always located in the same sentence. In this case, both methods fail to identify collocation candidates. Improving the preprocessor by adding a co-reference resolution engine should increase the overall numbers of collocation candidates and soften the consequences of the fact that some collocate types tend to be pronominalized more often than the others.

4.4. Typical argument extraction

Although some researches use (or assume the need of use of) syntactic information to extract typical arguments from the collocation lists or to filter them out, our study shows that both methods are suitable for the extraction of such verb-noun constructions. Both typical subjects and typical objects can be retrieved by either method, see examples below:

Typical subjects

(19) *арестовать* ('arrest')c10wc10

syntax: полиция ('police'), суд ('court')

window: год ('year'), полиция ('police'), суд ('court')

Typical objects

(20) *сломать* ('break')c10wc10

syntax: рука ('arm'), нога ('leg')

window: рука ('arm'), нога ('leg')

The possible way to take into consideration the particular type of arguments in the window-based method is to use the more granulated noun morphological features such as cases. However the distinguishing between these two cases or, in more complex cases, between subject and object collocates within the list of one verb was beyond the scope of our research.

5. Conclusion

In our study we have compared two methods of building collocation candidate lists within the framework of verb-noun collocation extraction. We have conducted an experiment on extracting and ranking collocation candidates from a large preprocessed corpus of news data using two different candidate extraction methods. An automatic comparison of collocation lists obtained using window-based and syntax-based candidate extractors has shown only a moderate level of correspondence. The detailed analysis of the comparison results makes it possible to identify common advantages and disadvantages of both methods.

In general, the window-based extractor seems to outperform the one based on a syntax-driven approach in terms of recall. Our results show that the simple syntax collocation model which only takes direct and prepositional verb-noun dependencies into account is not powerful enough. It is due to two basic phenomena. The first one is that there are a sufficient number of cases when the semantically related nouns are not immediate dependant of a verb. Moreover they can occur close to a verb but in another clause. The second one is the anaphora phenomena. The arguments of a verb can be pronominalized or omitted in real discourse especially as far as subject NPs is concerned. Adding special modules for syntax-based collocation extraction for treating these phenomena might improve the quality of the syntax-based method.

References

1. *Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N.* (2000), Dependency treebank for Russian: concept, tools, types of information. In Proceedings of the 18th conference on Computational linguistics — Volume 2, COLING '00, pp. 987–991.
2. *Breidt E.* (1993), Extraction of V-N-collocations from text corpora: A feasibility study for German, Proc. of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Columbus, USA.
3. *Church K. W., Hanks P.* (1990), Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), pp. 22–29.
4. *Evert S., Krenn B.* (2001), Methods for the qualitative evaluation of lexical association measures. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, pp. 188–195.
5. *Gildea D., Jurafsky D.* (2002), Automatic Labeling of Semantic Roles , *Computational Linguistics*, 28(3).
6. *Jagunova E., Pivovarova L.* (2010), The nature of collocations in Russian. The experience of automatic extraction and classification in news text. [‘Priroda kollokatsiy v russkom jazyke. Opyt avtomaticheskogo izvlechenia i klassifikatsii na material novostnykh tekstov’], *NTI*, 2, №6. M.
7. *Jongejan B., Dalianis H.* (2009), Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore : Association for Computational Linguistics, 2009. pp. 145–153.

8. *Khokhlova M.* (2008), Extracting Collocations in Russian : Statistics vs . Dictionary, JADT 2008: Proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12–14, 2008, pp. 613–624.
9. *Khokhlova M.* (2008), The experimental verification of collocation extraction [‘Eksperimentalnaja proverka metodov vydelenija kollokatsij’], *Slavica Helsingiensia*, 34. Helsinki. pp. 343–357.
10. *Khokhlova M.* (2009), Applying Word Sketches to Russian. In: Proceedings of Raslan 2009. Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, pp. 91–99.
11. *Kilgarriff A., Tugwell D.* (2002), Sketching words, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Marie-Hélène Corréard (Ed.) EURALEX, pp. 125–137.
12. *Klyshinskij E., Kochetkova N., Litvinov M., Maksimov V.* (2010), Automatic construction of word combination database using a huge text corpus [‘Avtomaticheskoje formirovanije bazy sochetaemosti slov na osnove ochen bolshogo korpusa tekstov’], *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2010”* [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2010”], Bekasovo, pp. 181–185.
13. *Kustova G., Toldova S.* (2009), RNC: Semantic filters for the verb disambiguation [‘NKRJA: semanticheskiye filtry dlja razreshenija mnogoznachnosti glagolov’]. Russian national corpus: 2006–2008. New results and perspectives. [‘Natsionalnyi korpus russkogo jazyka: 2006–2008. Novye rezultaty i perspektivy’]. Saint-Petersburg: Nestor-Istorija.
14. *Lin D.* (1998), Automatic Retrieval and Clustering of Similar Words, COLING-ACL98, Montreal, Canada.
15. *Nivre J., Hall J., Nilsson J.* (2006), Maltparser: A data-driven parser-generator for dependency parsing. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 2216–2219.
16. *Orliac B., Dillinger M.* (2003). Collocation extraction for machine translation, Proceedings of Machine Translation Summit IX, New Orleans, LA, USA, pp. 292–298.
17. *Pado S., Lapata M.* (2007), Dependency-based Construction of Semantic Space Models, *Computational Linguistics* 33(2), pp. 161–199.
18. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
19. *Sharoff S., Nivre J.* (2011), The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Proc. Dialogue 2011, Russian Conference on Computational Linguistics.
20. *Todirascu A., Gledhill C.* (2008), Extracting Collocations in Context: The case of Verb-Noun Constructions in English and Romanian. *Recherches Anglaises et Nord-Américaines (RANAM)*, Université Marc Bloch, Strasbourg.
21. *Todirascu A., Tufis D., Heid U., Gledhill C., Stefanescu D., Weller M., Rousselot F.* (2008), A Hybrid Approach to Extracting and Classifying Verb+Noun Constructions, Proceedings of LREC’2008, Marrakesh, Morocco.