# Method of POS-disambiguation Using Information about Words Co-occurrence (For Russian)

**Klyshinsky E.S.[1], Kochetkova N.A.[2], Litvinov M.I.[2], Maximov V.Yu.[1]**
[1]Keldysh IAM
Moscow, Russia, 125047 Miusskaya sq. 4
[2]Moscow State Institute of Electronics and Mathematics
Moscow, Russia, 109029 B. Tryokhsvyatitelsky s. 3
E-mail: klyshinsky@itas.miem.edu.ru, natalia_k_11@mail.ru, promithias@yandex.ru, vadimmax2000@mail.ru

**Abstract**

The article describes the complex method of part-of-speech disambiguation for texts in Russian. The introduced method is based on the information concerning the syntactic co-occurrence of Russian words. The article also discusses the method of building such corpus. This project is partially funded by RFBR grant 10-01-00805.

Keywords: learning corpora, words co-occurrence base, POS-disambiguation

## 1. Introduction

Part-of-speech disambiguation is an important problem in automatic text processing. At the time there exist many systems which solve this problem. The earliest projects use rule-based methods (see, for example, [Tapanainen and Voutilainen, 1994]). This approach is based on the following ideas: the system is supplied with some limiting rules which forbid or allow some certain words combinations. However, this method requires a time-consuming procedure of writing the rules. Besides, though these rules provide a good result, they often leave a considerable part of text not covered. In this connection there have appeared various statistical methods of automatic generation of such rules (for example [Brill, 1995]).

The n-gram method uses the statistical distribution of word combination in the text. Generally, n-gram model could be written down as follows:

$$P(w_i) = \arg\max P(w_i \mid w_{i-1}) * ... * P(w_i \mid w_{i-N}). \quad (1)$$

$P(w_i)$ is the probability of an unknown tag $<w_i>$ occurrence, if $<w_{i-N}>$ of the neighbours are known.

In order to avoid the problem of rare data and getting a zero probability for the occurrence of tag combination $<w_i \mid w_{i-1}, w_{i-2}>$, the smoothed probability can be applied for trigram model. The smoothed trigram model contains linear combinations of trigram, bigram and unigram probabilities:

$$Psmooth\ (w_i \mid w_{i-2} * w_{i-1}) = \lambda_3 * P(w_i \mid w_{i-2} * w_{i-1}) + \\ + \lambda_2 * P(w_i \mid w_{i-1}) + \lambda_1 * P(w_i) \quad (2)$$

where the sum of coefficients $\lambda_1+\lambda_2+\lambda_3 = 1$, $\lambda_1>0$, $\lambda_2>0$, $\lambda_3>0$. The values for $\lambda_1, \lambda_2, \lambda_3$ are obtained by solving the system of linear equations.

In [Zelenkov 2005] the authors in their disambiguation model had defined an unknown tag $w_i$ by involving not only the information on left neighbours, but also the right ones. We will use the similar approach when our system works with the trigram model. In this case the unknown tag is defined by involving the left neighbours $<w_{i-2}, w_{i-1}, w_i>$ (3), the right ones $<w_i, w_{i+1}, w_{i+2}>$ (4), and both the left and the right ones $<w_{i-1}, w_i, w_{i+1}>$ (5).

However, both the rule-based and trigram models require large tagged corpora of texts. The trigram rules which do not contain the information on a lexeme reflect specific language features, but the trigrams themselves (with lexemes inside) reflect rather the lexis in use. If the texts from another knowledge domain are given, the trigrams may show considerably worse results than for the initial corpus.

According to Google researches the digital collection of English texts they possess contains $10^{12}$ words. The British [BNC 2011] and America [ANC 2011] National Corpora contain about $10^8$ tagged words. According to the information on January, 2008 Russian National Corpus [RNC, 2011] contains about $5.8*10^6$ disambiguated words (and still remain). At present the process of filling up the latest corpora is rather frozen than active (unlike the situation for the first years of the project when it was being filled up intensively). The task of tagging (though automated) $10^{12}$ words, seems to be economically impracticable, and may be even unnecessary. The realization of practical applications for processing $10^9$ trigrams (the quantity estimation for English language could be found in [Google 2006]) will require a considerable amount of computational resources.

At present there are trigram bases accumulated that solve the problem with 94-95 % accuracy for Russian [Sokirko 2004]. The additional methods increases the quality of the disambiguation up to 97,3 % [Lyashevskaya 2010]. It is worthy to note that the application of rule-based methods requires essential time expenses. The application of trigrams demands a well-tagged corpus, and it is a costly problem too. The rule creating is also connected with a permanent work of linguists. The results of such work are never in vain, the output remains applicable to many other projects, but such results are helpless to improve the accuracy immediately. In this connection we had set a goal to develop a new method which would use results of the previous developments accumulated in this field and information from partial syntax analysis.

## 2. Obtaining statistical data on co-occurrence of words

It is widely acknowledged that a resolution of lexical ambiguity by means of a syntactic analysis allows to obtain high-quality results, although such approach requires a lot of resources. In this case it is recommended to apply other methods, for instance, n-grams. However, n-gram method requires a substantial preliminary work to prepare a tagged text corpus. We have decided to develop a disambiguation method, which uses the syntactic information (obtained in the automatic mode) without carrying out full syntactical parsing. In our researches we focused on Russian.

As the practice has shown, full parsing that would provide full constructing of the tree is not required to remove the most part of a homonymy (about 90%). As it happens, it is sufficient to include the rules of words collocation in nominal and verb phrases, folding of homogeneous parts of sentence, agreement of subject and predicate, prepositions and case government and some others, in total not exceeding 20 rules, which are described by context-free grammar. It is possible to have a more detailed look at the methods of formal description of language, for instance, in [Ermakov 2002].

To solve the problems mentioned above, it is necessary to create a method of getting information on a syntactic relationship for the words which are obtained from a non-tagged corpus. Preliminary experiments have shown that in Russian language approximately 50% of words appear to be part-of-speech unambiguous (up to 80% in conversation texts, in comparison with less than 40% for news in English). It means that there are no lexical homonyms for each of such words. So the probability to find a group of unambiguous words in a text is rather high.

The analysis of Russian sentence structure allows to determine some of its' syntactic characteristic features.

1) The noun phrase (NP) which follows the sole verb in the sentence is syntactically dependent on this verb.

2) The sole NP which opens the sentence and is followed by a verb, is syntactically subordinated to this verb.

3) The adjectives that are located before the first noun in the sentence, or between a verb and a noun, are syntactically subordinated to this noun.

4) The paragraphs 1-3 could be applied also to adverbial participles, and it is possible to consider participles instead of adjectives.

We had applied our method to the processing of several untagged corpora in Russian language. The total amount of these corpora included more than 4,2 billion of words. The text sources contain texts on various themes in Russian. The used corpora include the sources given in the Table 1.

The morphological tagging was made with the help of module of morphological analysis "Crosslator" developed by our team [Yolkeen 2003]. The volume of the databases obtained is listed in the table below. The numerator shows the detected total amount of unambiguous words with the given fixed type of syntactic relation. The denominator shows the amount of unique combinations of words of the given type.

| **Source** | **Amount mln w/u** | **Source** | **Amount mln w/u** |
|---|---|---|---|
| WebReading | 3049 | Lenta.ru | 33 |
| Moshkov's Library | 680 | Rossiyskaya Gazeta | 29 |
| RIA News | 156 | PCWeek RE | 28 |
| Fiction coll. | 120 | RBC | 21 |
| Nezavisimaya Gazeta | 89 | Compulenta. ru | 9 |
| | | **Total** | **4214** |

Table 1: Used corpora

| **Pair** | **Total, mln** | **>1, mln** | **>2, mln** |
|---|---|---|---|
| V+N | 243 / 10.89 | 237 / 5.27 | 235 / 4 |
| Ger+N | 40.8 / 2.76 | 39.3 / 1.25 | 38.7 / 0.91 |
| N+Adj | 67 / 2.15 | 66 / 1.13 | 65.6 / 0.9 |

Table 2: Obtained results

The analysis of the results (Table 2) has shown that the selected pairs contain 22200 verbs from 26400 represented in the morphological dictionary, 55200 nouns from 83000 and 27600 adjectives from 45300 represented in the dictionary. Such a significant amount of verbs could be explained by their low degree of ambiguity as compared with other parts of speech. A small number of adjectives could be explained by the fact that from several adjectives located immediately before a noun, only the first one was entered into the database. It should be noted that when the largest corpus had been integrated into the system, the number of lexemes has not been changed notably, but at the same time the number of pairs detected significantly increased. For example, the number of verbs has increased from 21500 up to 22200, whereas the number of unique combinations of verb + noun type has increased from 8,3 million to 10,9. Moreover, the amount of such combinations that had occurred more than 2 times, has increased from 2.3 to 4 million. Thus, it is possible to say that when a corpus contains more than one billion words, the lexis in use achieves its saturation limit, while its usage continues to change.

About 9 % of all word occurrences from the total amount of the corpus had been used to build a co-occurrence base. But even this percentage had appeared to be sufficient to construct a representative sample for a word co-occurrence statistics. The estimations have shown that the received word combinations contain not more than 3% of the errors mostly caused by an improper word order or neglect of some syntactically acceptable variants of collocations, deviances in projectivity and mistakes in the text. It is necessary to stress that all results had been obtained in the shortest terms without any

manual tagging of the corpus. Probably the results could be more representative, if we were to use some methods of part-of-speech disambiguation. However, the best methods give a 3-5 % error, and it would affect the accuracy of results but not noticeably. On the other hand, the sharp increase in corpus volume will allow to neglect the false alternatives at a higher level of occurrence and by these means preserve the quality level.

## 3. Complex Method of Disambiguation

After we had collected the co-occurrence base, which was sufficiently large, we have got all that was necessary to solve the main problem, that is, to create a method of disambiguation for texts in Russian on the basis of information on a syntactic co-occurrence of words.

Let us assume that in the sentence, which is being parsed, there are two words between which there are only several words or no words at all, and it is known that these two words could be linked by a syntactical relation. In this case, if we have other less probable variants of tagging these words, it is possible to assume that the variant with such link will be more probable. The most difficult thing here is to collect a sufficiently representative base of syntactic relations.

In this paper the rules shall be understood as an ordered set: $<\mathbf{v}_i, \mathbf{v}_{i+1}, \mathbf{v}_{i+2}>$, where $\mathbf{v}_i = <p_w, \{pr\}>$ is a short description of the word, $p_w$ is a part of speech of the word, and {pr} is a set of lexical parameters of the word. Thus, in such rules the lexemes of a word are not taken into account in contrast to the lexical characteristics of the word. A rule may be interpreted in different ways and can be written down as an occurrence $v_i$ with regard for its right neighbours, as an occurrence $v_{i+2}$ with regard for its left neighbours or as an occurrence $v_{i+1}$ with regard for its both neighbours. The set of rules has been obtained from the tagged corpus.

Following [Zelenkov 2005], we will make tagging of a word considering its right and left neighbours. In the mentioned above paper a tag of the word is defined only with regard for the nearest neighbours of current word. However, it is not necessary for such approach to produce the result that falls within the global maximum. The exhaustive search of word tagging variants is usually avoided, as it takes too much time.

As it already has been noted above, the ratio of unambiguous tokens is about 50% in Russian. In this connection there is always a sufficient probability to find a group of two unambiguous words. Moreover, the chance grows as the length of the sentence increases. If such groups are not found while searching a global maximum, the first word in the sentence will indirectly influence even the last word. In the case such groups are present, such relationship is cancelled, and the search of global criterion can be effected over the separate fragments of the sentence. It allows to increase essentially the speed of the algorithm. So the sentence “Так думал молодой повеса, / Летя в пыли на почтовых, / Всевышней волею Зевеса / Наследник всех своих родных.” (Such were a young rake's meditations – / By will of Zeus, the high and just, / The legatee of his relations – / As horses whirled him through the dust.) can be split into three independent parts: “Так думал молодой повеса, Летя”, “Летя в пыли на почтовых” and “Всевышней волею Зевеса Наследник всех своих родных”.

Thus, we no longer consider the problem $P_{sent} = \text{argmax}(\prod_{i=1}^{n_s} P(\mathbf{v}_i \mid \mathbf{v}_{i-1}, \mathbf{v}_{i-2}))$, where $n_s$ is a number of words in the sentence, but $P_{sent} = \prod_{i=1}^{n_f} \text{argmax}(\prod_{i=1}^{n_{fi}} P(\mathbf{v}_i \mid \mathbf{v}_{i-1}, \mathbf{v}_{i-2}))$, where $n_f$ is a number of fragments, $n_{fi}$ is a number of words in the i-*th* fragment. According to formulas (2)-(4), we consider both left and right neighbours of the word.

We seek the optimum from the edges of the fragment towards its center. It is obvious that product of the maximal values of probabilities for each word can give a global maximum. If this is not the case, but the values obtained from two sides had come to one and the same disambiguation of the word in the middle of the fragment, than we will also consider that we have a good enough solution. If variants of disambiguation of the word in the middle of the fragment are different for two solutions, the optimization is carried out for the accessible variants until they won't achieve one and the same decision. In any case, the optimization is not carried out even for an entire fragment, not mentioning the whole sentence.

The amount of unambiguous fragments can be increased by a preliminary disambiguation using another method. We use the described above base of syntactic dependences. So, let we have a set $\{<\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, p>\}$, $\mathbf{w}_i = <l_w, p_w, \{pr\}>$ is a complete description of the word where $l_w$ is a word’s lexeme, $w_1$ is a key word in the word-group (for example, a verb in the pair «verb+noun»), $w_2$ is a preposition (if any), $w_3$ is a dependent word, $p$ is a probability of word combination $\mathbf{w}_1 + \mathbf{w}_2 + \mathbf{w}_3$. In this case all rules are searched for every word of the sentence. It should be noticed that no word can participate in more than two rules. Thus, for each word it is necessary to calculate argmax $(p_1 + p_2)$, where $p_1$ and $p_2$ are the probabilities of rules containing this word in dominant and dependent position.

Actually, during the check of compatibility of the words among themselves, our system uses the following bigram model $P(w_i) = \text{argmax} P(w_i \mid w_{i-l})$, where $l$ means the distance (in number of words), at which the unknown word may stand from the known one. The rule containing the given word is selected in the following way. We take the floating window containing 10 words to the right and left. The dependent word must be located within this window, the preposition must be located before the dependent word, but there must be no main word between them. Besides, the adjective must lexically agree with a noun.

## 4. Results of experiments and discussion

As a result of our work we had obtain the Corpus of syntactical combinations of Russian words. The relations were achieved using untagged corpora of general lexis texts containing more than 4 billion words. The tagging was carried out “on the fly”. There had been revealed about 6 million of authentic unique word combinations which had occurred in the text more than 340 million times. According to our estimations, the amount of errors in the obtained corpora doesn't exceed 3 %. The number of word combination can be enlarged by processing the texts of a given new domain. Though, the investigations had shown that scientific texts use other constructions which reduce the amount of sampled combinations, for example, for speech and cognition verbs. Our method extracts about 9 % of tokens from common lexis texts. But news lines give us just about 5 %. Moreover, for scientific texts this number shortened to 3 %. So the method shows different productivity for different domains. Further experiments have discovered that the received results can be used for defining the style of texts.

So the suggested method allows almost automatically obtaining the information on word compatibility which further can be used, for instance, for parsing or at other stages of text processing. The method is also not strictly tied to the texts of a certain domain and has rather low cost of enlargement.

The estimation of the efficiency of the system with various parameters was carried out with carefully tokenized corpora that contained about 2300 words. Results were checked using Precision and Accuracy measures. The mere involving of information on word compatibility in Russian method had shown 71.98% Precision ratio and 96.75% Accuracy. This result is comparable with best results in selected area [Lee 2010]. The advantage of this method is in its` ability to be additionally adjusted to a new knowledge domain quickly and automatically (that is most important), in case a sufficiently large text corpus is available. The method gives an acceptable quality of disambiguation, unfortunately with not too large Precision.

The coverage ratio can be improved by application of trigram rules, which can be easily received, for example, from http://aot.ru, or by analysis of the tagged corpus in Russian (for example, http://ruscorpora.ru). The coverage ratio in this case has made 78%, but the accuracy has fallen to 95.6%. In [Sokirko 2004] it is mentioned that the systems Inxight and Trigram provide 94.5% and 94.6% accuracy accordingly, that is comparable with the results of our system. Further improvement of coverage ratio up to 81.3 % is possible in case of the improvement of optimal decision search algorithm which is described above, but it slightly brings down the accuracy. In the current state the method is not able to show an absolute coverage, because the part-of-speech list applied in this method was not full, it contained only the following: a verb, a verbal adverb, a participle, a noun, an adjective, a preposition and an adverb. Then, there was no information on some types of relations, for example, «noun+noun». Furthermore, the information on a compatibility of some words of Russian conceptually cannot be obtained because of fundamental homonymy of certain words. For example, the word "white" can be used as an adjective and as a noun in all its forms.

Our results are applicable to some (but not all) European languages. So the extremely unambiguous English doesn’t allow construct the words combinations database. Method can be applied for German or French but the rules should be completely rewritten. Problems like verbal detachable prefixes in German and reverse words order should be taken into account.

## References


Tapanainen P. and Voutilainen A. (1994): Tagging accurately - don‘t guess if you know. In Proceedings of the conference on applied natural language processing, 1994.

Brill E. (1995): Unsupervised learning of disambiguation rules for part of speech tagging. In Proceedings of the Third Workshop on Very Large Corpora, pages 1–13, 1995.

Zelenkov Yu.G, Segalovich Yu.A., Titov V.A. (2005): Statistical model of POS-disambiguation based on normalizing substitution and neighbor words positions // In Proc. of Annual International Conference "Dialogue’2005". [In Russian]

British National Corpus (2011): http://www.natcorp.ox.ac.uk/

American National Corpus (2011): http://americannationalcorpus.org/

Russian National Corpus (2011): http://www.ruscorpora.ru/

Google (2006): All Our N-gram are Belong to You // Google research blog, http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Sokirko A.V., Toldova S.Yu. (2004): Comparing a stochastic tagger based on Hidden Markov Model with a rule-based tagger for Russian // In Proc. of Corpus Linguistics – 2004, Saint-Petersburg. [In Russian]

Lyashevskaya O. at all (2010): NLP Evaluation: Russian Morphological Parsers // In Proc. of Annual International Conference "Dialogue’2010". [In Russian]

Ermakov A.E. (2002): Restricted syntactic analysis of text document in information retrieval systems // In Proc. of Annual International Conference "Dialogue’2002". [In Russian]

Yolkeen S.V., Klyshinsky E.S., Steklyannikov S.E. On problems of universal morpho-semantic dictionary // In Proc. of IEEE AIS’03 CAD-2003 – 2003 Divnomorskoe. [In Russian]

Lee Y.K., Haghighi A., Barzilay R. (2010) Simple Type-Level Unsupervised POS Tagging // In Proc. of EMNLP 2010