ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2011, Vol. 45, No. 1, pp. 15–19. © Allerton Press, Inc., 2011.
Original Russian Text © E.S. Klyshinskii, N.A. Kochetkova, M.I. Litvinov, V.Yu. Maksimov, 2011, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 2011, No. 1, pp. 31–35.

A Method for Disambiguation of Part-of-Speech Homonymy Based on Application of Syntactic Compatibility in the Russian Language

E. S. Klyshinskii, N. A. Kochetkova, M. I. Litvinov, and V. Yu. Maksimov

Received November 12, 2010

Abstract—This paper deals with the description of a complex method for the disambiguation of part-of-speech homonymy in Russian texts. The method is based on the data on syntactic compatibility of Russian words. A method for compiling a similar corpus is discussed.

Keywords: limiting rules, software model, syntactic compatibility, lexical disambiguation, word usage.

DOI: 10.3103/S0005105511010110

INTRODUCTION

Disambiguation of part-of-speech homonymy is one of the intrinsic problems in the automatic processing of texts. A large number of systems that aim to resolve the problem are available. The earliest works refer to methods based on rules (see, e.g., [1]). This approach is based on the fact that limiting rules are specified in a system to prohibit or fix certain word combinations. However, writing these rules is a time-consuming process. In addition, these rules, although they yield a good result, do not cover a major part of the text. In this connection, statistical methods for the automatic creation of similar rules have appeared [2].

The method of n-grams, which uses the statistics of the word combinations in a text, is an alternative to the above method. In the general form, an N-gram model is written as follows:

$$P(w_i) = \arg\max P(w_i|w_{i-1}) \times P(w_i|w_{i-2}) \times P(w_i|w_{i-N}). \quad (1)$$

This means that the probability of meeting an unknown tag is $\langle w_i \rangle$ if $\langle w_{i-N} \rangle$ neighbors are known.

For a tri-gram model a smoothened probability is used to avoid the problem of rare data and zero probability of the appearance of the tag combination $\langle w_i \mid w_{i-1} w_{i-2} w_{i-3} \rangle$. A smoothened tri-gram model contains linear combinations of tri-gram, bigram, and unigram probabilities:

$$\text{Psmooth}\,(w_i|w_{i-2} \times w_{i-1}) = \lambda_3 \times P(w_i|w_{i-2} \times w_{i-1}) + \lambda_2 \times P(w_i|w_{i-1}) + \lambda_1 \times P(w_i), \quad (2)$$

where the sum of the coefficients $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $\lambda_1 > 0$, $\lambda_2 > 0$, and $\lambda_3 > 0$. The values for λ_1, λ_2, and λ_3 are obtained by solving a set of linear equations.

The authors of [3] built the definition of an unknown tag w_i via not only the left neighbors but the right ones into their model of homonymy disambiguation. We use a similar approach when working with a system with an incorporated trigram model, i.e., an unknown tag is defined by its left neighbors $\langle w_{i-2}, w_{i-1}, w_i \rangle$ (3), by its right ones $\langle w_i, w_{i+1}, w_{i+2} \rangle$ (4), and by both left and right neighbors $\langle w_{i-1}, w_i, w_{i+1} \rangle$ (5).

One should note that trigrams do not necessarily go in succession. Thus, as has been reported elsewhere [4], dependant words and sentences are five to seven words apart from each other and the trigram model cannot take this dependence into account. The above-described presentation proposes a method for correcting this condition.

However, the rule-based model, like the trigram model, requires a large corpus of marked texts. Moreover, while trigram rules that don not contain any data on lexemes reflect language features, proper trigrams (with lexemes included) are more reflective of the vocabulary to be used. When changing the subject domain described in the texts, trigrams can yield considerably worse results than with using the corpus from which they were compiled.

According to research undertaken by the Google Company as of September 22, 2006, the digital collection of English texts that exists contains 10^{12} word combinations. the national corpora of Great Britain [5] and the USA [6] contain about 10^8 marked word combinations. According to the data as of January, 2008 (more than 2 years ago) the National Corpus of the Russian Language" [7] contained 5.8×10^6 word combinations with disambiguated homonymy. At the moment the corpus is frozen rather than increasing, as it was during the first years of its existence. Marking, even automatically, 10^{12} word combinations is a sophisticated, probably unnecessary, economic problem. The realization of practical applications with 10^9 trigrams (see the evaluation of the number for the English language in [8]) will need considerable computing resources.

Table 1.

Source	Amount, mln of word usages
WebReading.ru	3049
Moshkov's Library	680
RIA News	156
Additional fiction collection	120
Independent Newspaper	89
Lenta.ru	33
Russian Newspaper	29
PCWeek RE	28
RBC	21
Compulenta.ru	9
Total	**4214**

Despite the above said, at the moment the bases of trigram usage have been accumulated that allow solving the problem with an accuracy of about 94–95% [10].

It should be noted that application of such techniques is time-consuming to a large extent. The use of trigrams assumes the creation of a well-marked text corpus, which is a rather cost-based task. Creation of the rules also requires the permanent employment of linguists. Such an activity is not commonly wasted and can be used for other purposes; however, it does not allow improving results promptly. In this connection we posed a task in our work to create a new method that does not require a marked corpus of texts or use the groundwork that is available in this field.

STATISTICS OF CO-OCCURRING WORD USAGE

The authors have undertaken an attempt to create a method for lexical disambiguation that applies syntactic information without full syntactic analysis, with the syntactic data being retrieved in the automatic regime. The main attention was given to the Russian language in our research.

As practice shows, full syntactic analysis that provides full minimization of the parcel-tree is not required to disambiguate a major part of homonymy cases (about 90%). Inclusion of the rules of coordination in the nominal and verbal groups, minimization of homogeneous parts, coordination of a subject and a predicate, prepositional–case government, and several more rules, in total about 20, which are described by context-free grammar, turns out to be sufficient. The methods for the formal description of language are discussed in detail elsewhere [11].

To solve the above-mentioned problems it is necessary to create a method for retrieving the data on syntactic relations of a word derived from the unmarked corpus of a text. Preliminary experiments have shown about 30% of all word usages to be monosemantic, i.e., lexical homonyms are absent for each word. In this connection, the probability of encountering a group of monosemantic words is high.

Analysis of the sentence structure in the Russian language allows us to distinguish a number of syntactic features.

1. A nominal group following the only verb in a sentence is subordinate to this verb.

2. The only nominal group placed in the beginning of a sentence before the only verb is syntactically subordinate to the verb.

3. Adjectives located either before a noun that is the first in a sentence or between a noun and a verb are syntactically subordinate to this noun.

4. Sentences 1–3 are applicable to adverbial participles and participles can be considered instead of adjectives.

Further consideration makes it possible to distinguish other features to determine various word groups without full analysis of sentences. If a sufficiently large number of non-homonymic groups existed in the Russian language for which rules 1–4 were valid it would be possible to obtain the statistics of word co-occurrence. In the future statistics can be used, e.g., for lexical disambiguation. Let two words located not far from each other be present in a sentence in question and a syntactic relation can be set between them. In this case, with the occurrence of other less probable variants of word tagging, one may assume that the relation-forming variant is more probable. The main problem is to collect a representative base of syntactic relations of words.

According to the proposed technique, we processed several unmarked corpora of Russian texts. The total volume of the corpora was more than 4.2 billion word usages. The sources we used contain various texts in the Russian language. The composition of the corpora is given in Table 1.

The "Crosslator" morphological analysis module we developed was applied to morphological marking. The volume of the resulting bases is presented in Table 2.

The numerator shows the total number of detected non-homonymic word usages possessing a syntactic relation of a given type. The denominator shows the number of unique word combinations of a given type.

As the study results show, 22200 verbs of 26400 presented in the morphological dictionary, 55200 nouns of 83000, and 27600 adjectives of 45300 were involved in the distinguished pairs. A large number of verbs can be attributed to considerably rarer cases of homonymy. The low number of adjectives occurred because only the first adjective of several ones standing before a noun could be placed in the base. One should note that addition of the largest-volume corpus does not change the number of lexemes included in the result substantially; however, the number of pairs that occurred increased. Thus, for example, the number of verbs

Table 2.

Pair	Total occurrence, mln	Usage >1, mln	Usage >1, mln
Verb + noun	243/10.89	237/5.27	235/4
Gerund + noun	40.8/2.76	39.3/1.25	38.7/0.91
Noun + Adj.	67/2.15	66/1.13	65.6/0.9

increased from 21500 to 22200, whereas the number of unique word combinations verb + noun increased from 8.3 million to 10.9 million.

Therefore, it can be concluded that at the corpus volume more than a billion word combinations saturation occurs in the vocabulary while its usage continues changing.

To build the compatibility base totally about 1.5% of word usages of the collected corpus was used. However, even this quantity was sufficient to compile the representative statistics of word combinations. Estimation has shown that the phrases contain not more than 3% of the errors related to incorrect word orders, missing some syntactically permitted variations of combinations, violation of the projectivity, and errors in the text. The results would probably be more representative if we used the methods of disambiguation of part-of-speech homonymy. However, the best methods give an error of 3–5%, which would have an effect on the accuracy of the results. On the other hand, a sharp increase in the corpus volume allows cutting off variants at a higher level and thus retaining the quality level.

Upon having collected a sufficiently large base of word combinations we switched to the resolution of our main task, that is, the elaboration of the method for homonymy disambiguation in the Russian language using data on the lexical compatibility of words.

A COMPLEX METHOD FOR THE DISAMBIGUATION OF HOMONYMY

Within this work, the rules are taken to mean the ordered triple $\langle \mathbf{v}_i, \mathbf{v}_{i+1}, \mathbf{v}_{i+2} \rangle$, where $\mathbf{v}_i = \langle p_w, \{pr\} \rangle$ is a short description of a word, p_w is a part of speech of a word, and {pr} is a set of lexical parameters of a word. Therefore, the rule does not take into account the word lexeme although lexical characteristics of the word are considered. The rule can be interpreted in an arbitrary way and can be written as input of $\mathbf{v}_i$ accounting for its right neighbors, as an input of $\mathbf{v}_{i+2}$ accounting for its left neighbors, or $\mathbf{v}_{i+1}$ accounting for the input of its two neighbors from both sides. The set of rules was obtained by the marked corpus of texts.

In line with [3] we tag words accounting for their neighbors from both sides. The aforementioned work uses only the closest neighbors for tagging a word. However, this approach does not necessarily yield the result that enters into the global maximum. Examination of all possible tag options is not commonly used, since it is time consuming.

As mentioned above, about 30% of all word usages are non-homonymic in the Russian language. In this connection the probability of meeting a group of two non-homonymic words is rather high and increases with sentence length. In the absence of such groups in the search for global maximum the first word indirectly affects the last word. In the presence of such groups the relationship is broken and the global maximum can be searched using individual fragments of a sentence, which makes it possible to increase the rate of algorithm work. Figure 1 shows an example how a sentence is divided into fragments (punctuation is omitted).

Therefore, we go from the problem solution

$$P_{sent} = \operatorname{argmax}\left(\prod_{i=1}^{n_s} P(\mathbf{v}_i | \mathbf{v}_{i-1}, \mathbf{v}_{i-2})\right), \quad (6)$$

where n_s is the number of words in a sentence. Instead the criterion for a sentence as whole is formulated in the following way:

$$P_{sent} = \prod_{i=1}^{n_f} P_{fragm\ i}, \quad (7)$$

where $P_{fragm\ i} = \operatorname{argmax}\left(\prod_{i=1}^{n_{fi}} P(\mathbf{v}_i | \mathbf{v}_{i-1}, \mathbf{v}_{i-2})\right)$ is the probability of encountering the ith fragment of the sentence with the given set of tags, n_f is the number of fragments in the sentence, and n_{fi} is the number of words in the ith fragment. In this case not only the data on the right neighbors is used but the left as well according to formulae 2–4.

The global optimum is chosen from the ends of the fragment towards its center. Production of maximal values of the probabilities of each of the words will clearly give a global maximum. If this not so, but the values obtained from both sides came to the same result in the middle of the fragment, then we also believe that a global maximum was achieved. If in the middle of the sentence in joining fragments the variants diverge then optimization is carried out for the variants that are available up to the moment when they yield a single solution. In any case, optimization is not

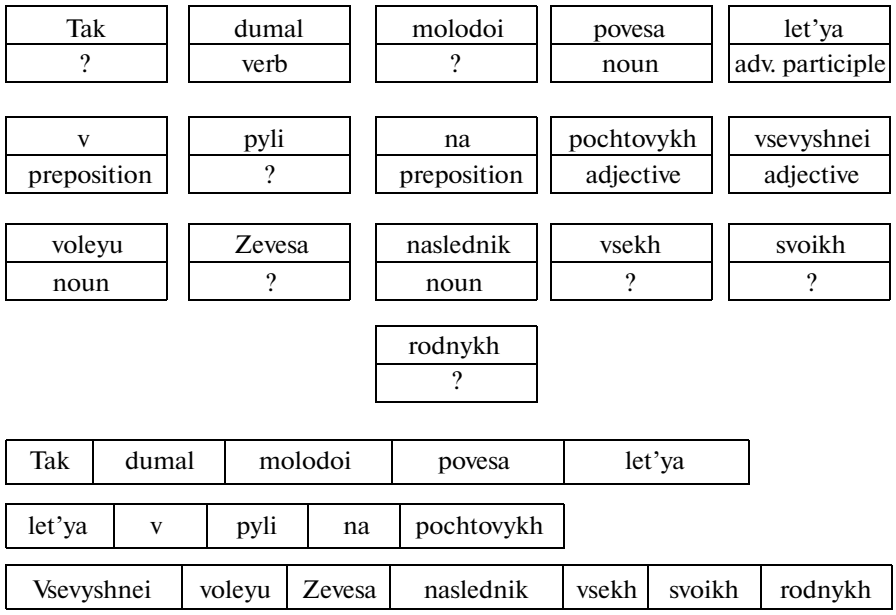

Fig. 1. An example of distinguishing fragments in a sentence.

carried out for the whole fragment not speaking about the sentence optimization

Thus, let us have the set $\{\langle \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, p\rangle\}$. Here $w_i = \langle l_w, p_w, \{pr\}\rangle$ is the full transcription of a word wherein l_w is a word lexeme, $\mathbf{w}_1$ is the leading word in a word combination (e.g., the verb in the pair "verb + noun" or the noun in the pair "noun + adjective"), $\mathbf{w}_2$ is a preposition (if available), $\mathbf{w}_3$ is a subordinate word, p is the probability to encounter the similar phrase. Then for each word we search for rules in which it is involved. Therefore, for each word it is necessary to calculate $\text{argmax}(p_1 + p_2)$, where p_1 and p_2 are probabilities of the rules in which the given word takes the main and subordinate positions.

Actually, in checking the words for compatibility with each other the bigram model is used or the system

$$P(w_i) = \text{argmax}\, P(w_i \mid w_{i-1}), \qquad (8)$$

where l is the distance at which an unknown word can stand from a known. Commonly l varies within 5–7 and allows taking the long-range relations in the sentence into account.

In our case the rule that involves the given word is chosen as follows. A window 7-word long from the current word is taken both to the left and to the right. The subordinate word should be found in this window and the preposition should precede the subordinate word but the main word should not stand between them. Moreover, the adjective should agree with the noun.

EXPERIMENTAL RESULTS

The results of system operation with different parameters were evaluated along a thoroughly marked track containing about 2300 word usages. Two parameters of the system's operational quality were evaluated by analogy with the information retrieval domain, namely, precision (the percentage of true answers from all those given by the system) and accuracy (the percentage of true answers from the entire track proposed).

$\text{Precision} = A_t/(A_t + A_{fa})$

$\text{Accuracy} = A_t/(A_t + A_{fa} + A_{fn})$

A_t is the number of true answers given.

A_{fa} is the number of false answers given.

A_{fn} is the number of answers that were not given.

During disambiguation of part-of-speech homonymy using the data on word compatibility in the Russian language, values of 71.98% for precision and 96.75% for accuracy were obtained. The advantage of the method is that it can be adjusted rapidly and, most importantly, automatically to the chosen subject domain if a sufficient–volume corpus of text documents is available. The method provides a reasonable accuracy of homonymy disambiguation, although the precision is not sufficiently high.

The precision parameter can be improved via application of trigram rules that are obtained easily from, e.g., the Internet resource http://www.aot.ru, or via analysis of the marked corpus in the Russian language (e.g. http://www.ruscorpora.ru). The precision in this case is 78%, yet the accuracy decreased to 95.6%. As was mentioned elsewhere [9], the Inxight and Trigram systems give an accuracy of 94.5% and

94.6%, respectively, which is comparable with the results of our system operation. With the improved algorithm for optimal solution search described above further improvement precision up to 81.3% is possible; however, it is accompanied by deterioration of the accuracy.

DISCUSSION

We obtained a corpus of syntactic relations between words of the Russian language. The relations were traced by the unmarked corpus of texts of a general vocabulary consisting of more than 4 billion words. The corpus was marked on the fly. In total 6 million reliable unique word combinations were obtained that were encountered in the texts more than 340 million times. By our evaluation, the number of errors did not exceed 2% in the obtained corpus.

The base of word compatibility can be added from the texts of a chosen subject domain; however, studies show that scientific texts use different constructions that reduce the number of distinguished word combinations, e.g., when we deal with speech and mental activity verbs. Thus, about 9% of the word combinations are used in literary texts whereas news highlights contain about 5% and scientific texts contain about 3% of the word combinations.

The method we suggest makes it possible to obtain the data on word compatibility that can be used in future e.g., in syntactic analysis or at some other stages of text processing, almost totally automatically.

The obtained corpus of word compatibility was used to create the method for disambiguation of part-of-speech homonymy. The method shows a good accuracy of up to 96.75%. However, its precision is not very good, viz., up to 81.3% with the use of trigram rules. In its pure form the method is unable to achieve 100% precision, since it uses a limited list of parts of speech, namely, verbs, adverbial participles, participles, nouns, adjectives, prepositions, and adverbs. Also the data is absent about some kinds of relationships, e.g., "noun + noun." In addition, the data on the compatibility of certain words in the Russian language cannot be obtained in principle because of the full homonymy of some words, e.g., "belyi" as an adjective and a noun.

The absence of strong connection to texts of a specified topic and the cheap supplementation procedure can be considered as advantages of the method.

It is planned to preliminarily disambiguate homonymy in order to increase the number of word combinations in a corpus. As a result, the reliability of the word combinations that are available would increase significantly. Moreover, we plan to improve the precision of homonymy disambiguation by adding "noun + noun" combination.

ACKNOWLEDGMENTS

The authors are grateful to E.V. Yagunova and L.M. Pivovarskaya for analysis of these findings. This work was supported in part by the Federal Targeted Program Scientific and Pedagogical Staff of Innovative Russia for 2009–2013.

REFERENCES

1. Tapanainen, P. and Voutilainen, A., Tagging Accurately—Don't Guess If You Know, in *Proc.Conf. on Applied Natural Language Processing*, 1994.
2. Brill, E., Unsupervised Learning of Disambiguous Rules for Part-of-Speech Tagging, in *Proc. 3rd Workshop on Very Large Corpora*, 1995, pp. 1–13.
3. Zelenkov, Yu.G., Segalovich, Yu.A., and Titov, V.A., Probabilistic Model of Disambiguation of Morphological Homonymy Based on Normalizing Substitutions and Positions of Neighboring Words, *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: trudy Mezhdunarodnogo seminara "Dialog'2005* (Computational Linguistics and Intelligent Technologies: Proc. Int. Workshop "Dialogue'2005",) 2005.
4. Protasov, S.V., Derivation and Evaluation of Parameters of the Long-Range Tri-Gram Model of Language, *Proc. Int. Workshop "Dialogue'2005" on Computational Linguistics and Intelligent Technologies*, 2005.
5. *Natsional'nyi korpus Velikobritanii* (National Corpus of Great Britain), http://www.natcorp.ox.ac.uk/.
6. *Natsional'nyi korpus SShA* (National Corpus of the USA), http://americanannationalcorpus.org/.
7. *Natsional'nyi korpus russkogo yazyka* (National Corpus of the Russian Language), http://www.ruscorpora.ru.
8. *All Our N-Grams Are Belong to You*, http://googlesearch.blogpost.com/2006/08/all-our-n-graml-are-belong-to-you.html.
9. Sokirko, A.V. and Toldova, S.Yu., Comparison of the Efficacy of Two Techniques for Lexical and Morphological Disambiguation in the Russian Language (A Hidden Markov Model and Syntactical Analyzer of Noun Groups), *Mezdunarodnaya konferentsiya "Korpusnaya lingvistika 2004"* (Int. Conf. "Corpora Linguistics 2004"), S.-Peterburg, 2004.
10. Lyashevskaya, O.N. et al., Evaluation of Methods of Automatic Analysis of Texts: Morphological Parsers in the Russian Language, *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: trudy Mezhdunarodnogo seminara "Dialog'2010"* (Computational Linguistics and Intelligent Technologies: Proc. Int. Workshop "Dialogue'2010"), 2010.
11. Ermakov, A.E., Incomplete Text Analysis in Information-Retrieval Systems, *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: trudy Mezhdunarodnogo seminara "Dialog'2002"* (Computational Linguistics and Intelligent Technologies: Proc. Int. Workshop "Dialogue'2002), Moscow: Nauka, 2002, vol. 2.