APPLIED PROBLEMS OF FUNCTIONAL HOMONYMY RESOLUTION FOR RUSSIAN LANGUAGE

Olga Nevzorova, Julia Zin'kina, Nicolaj Pjatkin

Abstract: Applied problems of functional homonymy resolution for Russian language are investigated in the work. The results obtained while using the method of functional homonymy resolution based on contextual rules are presented. Structural characteristics of minimal contextual rules for different types of functional homonymy are researched. Particular attention is paid to studying the control structure of the rules, which allows for the homonymy resolution accuracy not less than 95%. The contextual rules constructed have been realized in the system of technical text analysis.

Keyword: natural language processing, functional homonymy, resolution of homonymy

ACM Classification Keywords: H.3.1.Information storage and retrieval: linguistic processing

Introduction

There is no common opinion on the phenomenon of homonymy in linguistic literature. The content of the phenomenon, principles of classification, classificatory schemes are being discussed. The most common classification divides homonyms into lexical ones, i.e. referring to the same part of speech, and grammatical ones, i.e. referring to different parts of speech. A term "functional homonyms" is also spread in linguistic literature. This term, offered by O.S. Ahmanova, has been accepted in the works by V.V. Babajceva [Babajceva, 1967] and her followers. V.V. Babajceva defines functional homonyms as "words with homophony, historically coming from one word family, referring to different parts of speech" and spreads the occurrence of functional homonymy not only on autosemantic parts of speech, but also on synsemantic ones. In the actual work methods of automatical resolution of functional homonymy of different types are researched.

Success of applied research in computer linguistics depends remarkably on the availability of appropriate linguistic resources, lexicographical ones being the most important. In the recent years dictionaries of homonyms of Russian language by different authors have been published. In these dictionaries, the phenomenon of homonymy has been represented with various degree of fullness. Thus, for example, in N.P. Kolesnikov's dictionary [Kolesnikov, 1978] the phenomenon of homonymy is understood in an extended sense, homoforms, homophones and homographs being included into the circle of the occurrences studied besides lexical homonyms. Attempts on describing functional homonyms have been made in separate homonym dictionaries by O.S. Ahmanova [Ahmanova, 1984], O.M. Kim [Kim et al., 2004]. The appearance of an Internet-resource by N.G. Anoshkina [Anoshkina, 2001], attempting to gather all grammatical homonyms, stimulated further development of theoretical and applied research, including that on the classification of functional homonymy types. The grand problem is mismatch of grammatical descriptions of homonyms in these dictionaries. For example, the comparison of the grammatical descriptions of 560 homonyms terminating on letter 'o' in [1-4] have shown that only three homonyms have been described with the same grammatical features.

In the work [Kobzareva et al., 2002] a classification of 58 homonymy types is given. This classification served as a base for working out rules of contextual resolution of functional homonymy. The rules are offered in the article. In the course of research some changes have been made to the basic classification of functional homonymy types. The changes were connected with the appearance of new subtypes, addition and exclusion of some types. It is obvious that it needs to develop new dictionary of functional homonyms on the basis of representative corpus of Russian texts for applied research.

Method of Contextual Resolution of Functional Homonymy (Problems of Building up Contextual Rules)

Theoretical research on the problem of functional homonymy resolution in texts has a long history. At the end of the 50-s in works by K.E. Harper [Harper, 1956], A. Caplan [Caplan, 1955], studying and describing contextual conditions in which some or another meaning of the word would be realized was accepted as the main way of homonymy resolution. Either the surroundings of the word in the text or the words with which the word given was used would be implied under "context". The question of minimal resolving context was also actual for the researches. Noteworthy are the results obtained by A. Caplan [Caplan, 1955] in his research on minimal resolving context. About 140 frequently used polysemantic English words (lexical homonyms mainly) in different contextual surroundings have been analyzed in the work.

The following types of contexts have been selected:

1). Combination with the preceding word – P1.

2). Combination with the following word – F1.

3). Combination with both preceding word and following word – B1.

4). Combination with two preceding words – P2.

5). Combination with two following words – F2.

6). Combination with two preceding words and two following words – B2.

7). The whole sentence – S.

The main conclusion implied that B1 chain was more productive as regards the effect of reducing polysemy (the proportion of the quantity of word meanings in concrete context to their number in zero context) than P2 and F2 contexts and would be almost equal to the effect given by S. Another conclusion emphasized the importance of material context type. That is, whether autosemantic parts of speech or so-called "particles" (including prepositions, conjunctions, auxiliary verbs, articles, pronouns, adverbs like *there* etc) exist in the direct surroundings of the word. The context with autosemantic parts of speech gives much better results than that with "particles". General conclusions by A. Caplan imply that the most useful context is the one consisting of one word to the left and one word to the right from the word given. If one word of the context is a particle, the context should be extended to two words on both sides.

Despite numerous references to the results quoted above in Western literature of the 60-s, their practical usage for Russian language in real contexts is hardly possible. The real situation with the resolution of functional and lexical homonymy in Russian language is far more complicated and cannot be resolved on the basis of simplified rules. In the work [Kobzareva et al., 2002], a special dictionary of diagnostic situations (DDS) has been developed. These situations assign linear structure description of the minimal context necessary for the identification of the homonym's part of speech.

DDS situations consist of two parts:

- component chain of the sentence (may be discontinuous) marks the situation;

- conditions put on the markers and their surroundings, defining the meaning of the homonym.

While analyzing the examples given in the article [Kobzareva et al., 2002], it is remarkable that the borders of the minimal resolving context become moveable. Symbols belonging to a certain multitude (varying for different homonymy types) act as borders. Besides, the authors allow for discontinuous contexts. This complicates the situation even further. Besides, yet another thing should be pointed out. With a set of rules used for the resolution of some homonymy type one of the main problems is building up a control structure for the order of the usage.

Saying it another way, the method of contextual resolution of functional homonymy offered by the authors of the present article includes:

- Establishing a full classification of functional homonymy types.

- Selecting the minimal set of resolving contexts for each type.

The minimality of the set means that for every functional homonymy type it should be estimated, how difficult it would be to recognize each part of speech belonging to the type. Then the set of resolving contexts (SRC) with minimal difficulty of resolution should be built up. The algorithmic form of this demand will look as follows:

If a rule from SRC has been applied to a functional homonym of T1 or T2 type, then the type of the homonym is defined by the rule applied, else the alternative type is given to X.

- Building up a control structure for the SRC, allowing for the maximum resolution accuracy.

Classification of the Functional Homonymy Types

The ground for basic classification of functional homonymy types was the classification offered in the article [Kobzareva et al., 2002]. It had been built up on the basis of N.G. Anoshkina's dictionary of grammatical homonyms [Anoshkina, 2001]. The classification proposes 58 types of homonyms, the first ten types being the most numerous (the whole number of homonyms is 2965). Less than 5 homonyms are included into each of 26 other types. In spite of the undoubted importance of the classification, it requires further development. The latter concerns the addition of new types as well as additional subdivision of types and creation of subtypes. For example, a new type "short form of adjective # category of state" like *gotovo (*in Russian*)/ ready*. Second kind of expansions is connected, for instance, with three subtypes in Vf/N* type, where Vf stands for a verb form and N* = {N - a noun, Npr - pronominal noun}.

The example in Russian is:

1) bereg — N - *a brink*

2) bereg (inf. berech') — V, Past Tense - *saved*

The subtypes have been selected on the basis of a syntactic criterium, i.e. the recognition of their representatives requires working out remarkably different rules and control structures (the examples are in Russian):

<Vf/N*>$_1$: *pojmu* = 1) N - *bottom-land*; 2) V, Future Tense - *I will understand*;
<Vf/N*>$_2$: *l'nu* = 1) N – *flax*; 2) V, Present Tense - *I am clinging to sth*;
<Vf/N*>$_{3.1}$: *stali* = 1) V, Past Tense - *we became*; 2) N - *steel*;
<Vf/N*>$_{3.2}$: *zarosli* = 1) V, Past Tense - *became overgrown*; 2) N – *bush.*

Rules of Resolving Functional Homonymy of Some Types (Realization)

In this part, an example of resolving functional homonymy of type N*/Comp, here N* ={N =a noun, Npr=pronominal noun}, Comp – comparative. Let the following system of denotations be introduced:

X – functional homonym; P– preposition. The term $X_{\bigcap_{pgn} N^*}$ means that X is complied with N* by the grammatical

characteristics given (p – case, g – gender, n – singular/plural).

A Z may be present in the rule. It means that an inserted construction of some type may be present here. In Russian language, the discontinuance of a linear sequence because of such constructions is possible almost everywhere in the sentence. We have used a typology of such constructions allowing to identify them. For example, constructions expressing feelings and emotions – '*k schast'ju' (in Russian)/ fortunately* –, one expressing the degree of trustworthiness – '*nesomnenno' (in Russian)/ assuredly* – etc. The recognition of them is a major point, as punctuation marks belong to the multitude of context restrainers. Consequently, recognizing the function of the punctuation mark is important. That is, we should know the one separating an inserted construction from all the others, as it is not considered a context constraint.

While working out the rules of resolution of functional homonymy of different types we have offered a method of multiplex resolution of functional homonymy. It allows to resolve groups of homogeneous homonyms. These rules are particularly typical for some certain types of homonymy, like N*/A* type (A*={A=full adjective form, Av=participle, Apr=pronominal adjective}), D/Abr type, D/Abr/Vsp type (D – adverb, Abr - the short form of the adjective, Vsp - predicate noun) and others.

Let us exemplify the group of contextual rules for N*/Comp type. The minimal set of resolving contexts was selected for resolving an homonym as Comp. The control structure of the general rule defines the order of usage and the results (in rectangles on the scheme).

$$1)\quad if \quad [\frac{P \cap_X (Z)\overline{N} \cap_p}{\leq 3} X\,] \quad then \quad X = N*$$

$$2)\quad if \quad [\frac{A* \cap_{pgn} x(Z)\overline{N} \cap_{A*}}{\leq 3} X\,] \quad then \quad X = N*$$

$$3)\quad if \quad [\frac{X(Z)N*_{p2}}{\leq 5}] / [\frac{X(Z)\bar{,}A*_{p2}}{\leq 4}] \quad then \quad X = Comp$$

$$4)\quad if \quad [\frac{X\ ,than}{\leq 4}] / [\frac{X\ ,that}{\leq 4}] / [\frac{that\ X}{\leq 3}] \quad then \quad X = Comp$$

$$5)\quad if \quad [\frac{V_{SV}(Z)X}{\leq 2/\otimes}] \quad then \quad X = Comp$$

$$6)\quad if \quad [\frac{all\ X}{\leq 2/\otimes}] \quad then \quad X = Comp$$

$$7)\quad if \quad [Comp\ \exists < S_{\&} > X\,] / [X\ \exists < S_{\&} > Comp\,] \quad then \quad X = Comp$$

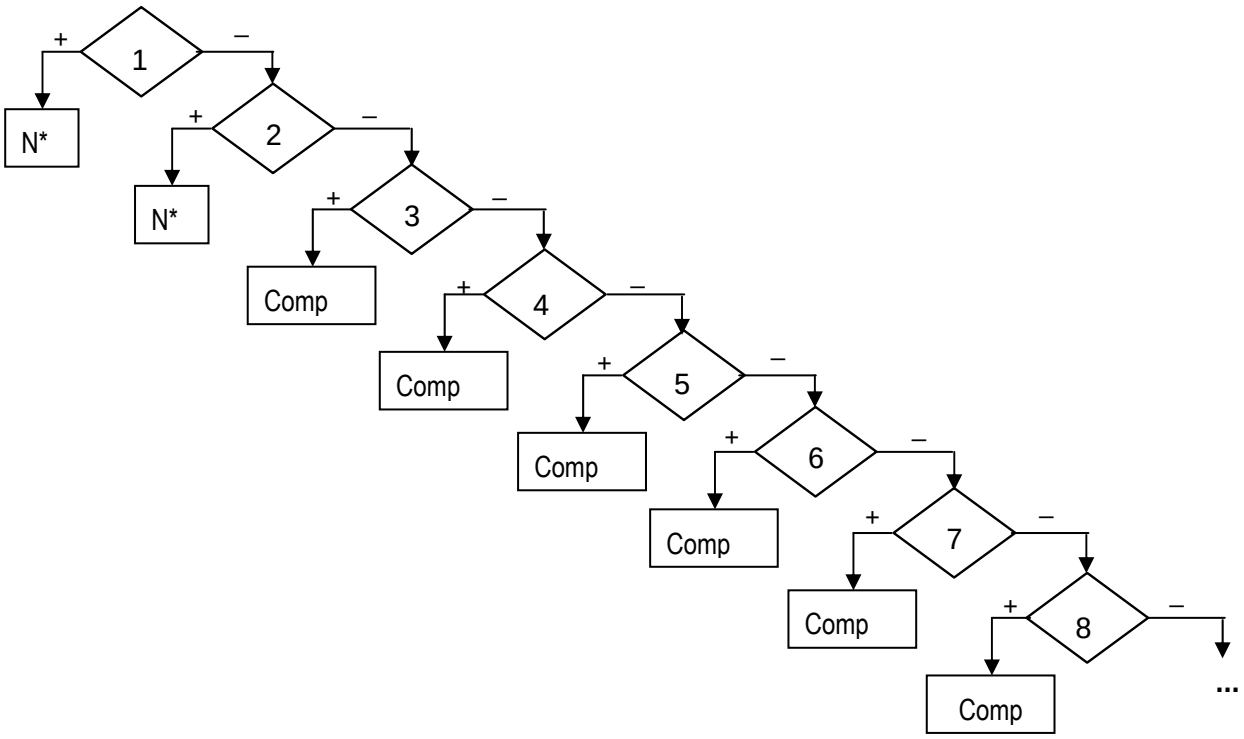
$$8)\quad if \quad [\frac{D_{md}(Z)X}{\leq 2/\otimes}] \quad then \quad X = Comp$$

$$9)\quad if \quad length(\otimes\frac{X}{\leq 4}\otimes) \quad then \quad X = Comp$$

$$10)\quad if \quad [\frac{V_f(Z)X}{\leq 2/\otimes}] / [\frac{X(Z)V_f}{\leq 2/\otimes}] \quad then \quad X = Comp$$

$$else \quad X = N*$$

Picture 1. Rules of resolution of functional homonymy of N*/Comptype



Picture 2. The control structure of the general N*/Comp rule

Conclusion

Program realization of syntactical processing module of technical texts analysis system is currently being completed. The module is to comprise the process of resolution of functional homonymy. Real technical texts, however, contain not all of the types. There are no types with interjections, for example, or with stylistically substandard words. Other limitations are connected with the vocabulary of technical texts itself (there are very few imperative verb forms, for instance). Nevertheless, contextual rules allow for such verbs to be recognized. Module setting operations allow for the amount of homonymy types liable to resolution to be changed.

Testing of the program module for the resolution of functional homonymy has given good results on the types realized. For some types, the accuracy of resolution is 100%, in the worst cases it is not less than 95%. The reasons for erroneous situations' appearance are accidental concord in the context analyzed, context insufficiency or resource insufficiency (absence of a case frame dictionary for different parts of speech). Some mistakes in the resolution can be sorted out in the course of further analysis.

Acknowledgements

The work has been completed with partial support of Russian Foundation Basic Research (grant № 05-07-90257).

Bibliography

[Ahmanova, 1984] A.S. Ahmanova. Slovar omonimov russkogo jazyka. M., 1984. (In Russian).

[Kolesnikov, 1978] N.P. Kolesnikov. Slovar omonimov russkogo jazyka. Tbilisi, 1978. (In Russian)

[Anoshkina, 2001] J.G. Anoshkina. Slovar omonimichnyh slovoform russkogo jazyka. M: Mashinnyj fond russkogo jazyka Instituta russkogo jazyka RAN, 2001. (In Russian) (http://irlras-cfrl.rema.ru:8100/homoforms/index.htm).

[Caplan, 1955] A. Caplan. An Experimental Study of Ambiguity and Context // Mech. Translation, vol. 2, No 2, Nov. 1955.

[Harper, 1956] K.E. Harper. Contextual Analysis // Mech. Translation, vol. 4, No 3, Dec. 1956.

[Kobzareva et al., 2002] T.U. Kobzareva, R.N. Afanasiev. Universalnyj modul predsintaksicheskogo analiza omonimii chastej rechi b RJa na osnove slovarja diagnosticheskih situacij // Trudy mezhdunar. konferencii Dialog'2002. M., 2002. S. 258-268. (In Russian).

[Kobzareva et al., 2002] O.M. Kim, I.E. Ostrovkina. Slovar graamaticheskih omonimov russkogo jazyka. M., 2004. (In Russian).

[Babajceva, 1967] V.V. Babajceva. Perehodnye konstrukcii v sintaksise. Voronezh, 1967. (In Russian).

Authors' Information

Olga Nevzorova - Research Institute of Mathematics and Mechanics, Kazan State Pedagogical University, Kazan, Russia; e-mail: olga.Nevzorova@ksu.ru

Julia Zin'kina - Kazan State University

Nicolaj Pjatkin - Research Institute of Mathematics and Mechanics, Kazan, Russia; e-mail: nikolaip@mail.ru