GRAMMATICAL DISAMBIGUATION IN THE TATAR LANGUAGE CORPUS

Bulat Khakimov
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan Federal University, Kazan, Russia
khakeem@yandex.ru

Rinat Gilmullin
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan Federal University, Kazan, Russia
rinatgilmullin@gmail.com

Ramil Gataullin
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan Federal University, Kazan, Russia
ramil.gata@gmail.com

ABSTRACT

This article concerns the issues of corpus-oriented study of the most frequent types of grammatical homonymy in the Tatar language and the possiblities for automation of the disambiguation process in the corpus. The authors determine the relevance of alternative parses generated in the process of automatic morphological analysis in terms of real linguistic ambiguity. This work presents a variant of classification of frequent homoforms and methods for their disambiguation, and it estimates the potential impact on the corpus.

Keywords: linguistic corpus, Tatar language, grammatical homonymy, homoform, disambiguation

1 Introduction

The problem of grammatical ambiguity and its resolution is one of the most pressing problems in modern computer and corpus linguistics [1]. “Tugan Tel” Tatar National Corpus, that was developed in the “Applied semiotics” Research Institute of the Tatarstan Academy of Sciences [2], employs the system of automatic morphological annotation on the basis of our own morphological analyzer [3]. In order to adequately reflect the specifics of the Tatar language, a morphological standard of the corpus was developed [4]. Research on specification and improvement of the metalanguage for the description of a Tatar wordform is currently carried out [5]. The general conception of the corpus is presented in [6]. To implement the grammatical disambiguation in the Tatar National Corpus, developers have conducted a study of contextual constraints of different types of grammatical homonyms, involving statistical corpus data, and suggest the methods of automatic grammatical disambiguation for the Tatar language.

2 Statistical Characteristics of the Corpus

At the initial stage of work we obtained the statistical data on the frequency of wordforms with alternative parses, presented in Table 1, from the database of texts of the Tatar National Corpus [2]. The total volume of the corpus is 21,940,452 word usages, the proportion of Word usages with alternative parses is 25.75%.

№	Alternative parses	Amount	Proportion in the corpus
1	Wordforms with alternative	5650820	25,75%

	parses		
	of which:		
2	2 parses	4282108	19,51%
3	3 parses	1045392	4,76%
4	4 parses	296547	1,35%
5	5 and more parses	26773	0,12%
6	Wordforms with alternative parses in the sample	21940452	100%

Table-1. Some statistical charasteristics of Corpus

To identify the most frequent types of homonymy in the corpus and to assess their relevance in terms of real language homonymy, a sample of 500 most frequent combinations of alternative parses was created. On its basis 150 types with two parsing options were selected for further analysis, because this parsing type is presented in the corpus in the biggest proportion.

3 Relevance Evaluation of Types of Homonyms

In the first phase of work, irrelevant combinations of homonyms were identified. In such combinations alternative parses often appear because of the errors of the morpholoical analyzer, that is due to the redundancy in the stem set or in the model of inflection. Some cases are caused by incorrect morphonological rules of the analyzer; correction of these rules also allows to exclude the cases of ambiguity belonging to the specified types.

The cases conditioned by the disuse of one of the parsing options present special interest. We refer to such cases as irrelevant, because the potential wordforms, which are automatically generated during the work of the morphological analyzer, are not represented in the actual speech use. A corresponding set of wordforms was experimentally determined for them.

The suggested measures on the exclusion of irrelevant types of homonymy have reduced the number of homonymous parses in the corpus by about 8.5% (2.1% of the total volume of texts in the corpus).

4 Most Frequent Types of Grammatical Homonymy

For the most frequent linguistically relevant types of homonyms we have made a classification, which groups separate automatically determined subtypes. The following frequent types of homonyms were singled out:

1. Noun vs Pronoun
2. Verb vs Noun/Adjective
3. Pronoun vs Numeral
4. Noun vs Adjective
5. Postposition vs Noun/Numeral
6. Noun vs Adverb
7. Adjective vs Noun with attributive affix
8. Noun/Adjective vs Noun with possessive affix
9. Adjective vs Noun in aditive case
10. Adjective vs Verb
11. Verb vs Verb
12. Adjective vs Adverb
13. Pronoun vs Pronoun in locative-temporal case
14. Noun vs Adjective with affix -chA
15. Pronoun vs Noun

All types except type 1, 3, 5, 6, 9 and 15, are represented by a set of regularly formed wordforms, which possess a certain number of grammatical features. Contextual disambiguation rules for these types are conditioned by these characteristics and the characteristics of the disambiguating context.

Type 1 is represented by a single frequent word *ul* (‘he/son’). Different context principles work for each of the part-of-speech alternatives.

Type 3 is also represented by only one frequent word *ber*, which is used both in the meaning of the numeral 'one' and in the function of the indefinite pronoun, that is close to the function of the indefinite article. Each part-of-speech alternative has its own context patterns.

Type 5 includes four subtypes. Each of them is represented by one word – postposition: *öçen* ('for'), *turında* ('about') and *buyınça* ('on'), or pronoun: *tege* ('that'). Each of these words has a homonym, which is a noun in a definite form. Grammatical characteristics of homonymous words and syntactic functions of the respective postpositions define context rules for this type.

Types 6 and 9 are represented by the lexemes *bik* ('very/bolt') and *başka* ('other/head+ DIR'), respectively.

Type 15 is also an example of one wordform homonymy; it is represented by the word *bez* ('we/awl').

The total number of all types of word usages is 1624839. The proportion in the corpus sample is 7,4% (21940452 word usages). The proportion among the homonymous parses is 28,7% (5650820 in the indicated corpus sample).

This variant of classification does not include another special case of verb forms homonymy, which is related to the multifunctionality of voice affixes. Thus, a statistical study of corpus data has shown that the total number of such cases of homonymy in the analyzed sample of texts is 408346 word usages (1.8% of the total volume of texts and 7.2% of all the alternative parses). The most frequent subtype among them is the V - V + REFL subtype, where one and the same verbal form can stand both for a separate lexeme, which is included on its own in the stem set, and the voice form of another lexeme. For example, *ezlänergä, totınırga, yaşerenergä, seltänergä, ağulanırğa, alınırğa*. Disambiguation of this type is not a trivial task, and in many cases requires consideration of not only morpho-syntactic, but also semantic characteristics of the disambiguating context.

5 Context Rules for Automatic Grammatical Disambiguation

In order to make use of classical methods of grammatical disambiguation based on context rules, we classified the types of homonyms, of which homoforms represent the biggest part. The full classification of types of homonyms (analysis of the full range of types) is an extremely time-consuming and pragmatically unreasonable task, as the Tatar language belongs to the agglutinative languages, where the number of morphemes that can be attached to the stem is theoretically unlimited. For example, in the above mentioned corpus of Tatar texts, which includes more than 21 million Word usages, there are more that 7000 types of homoforms.

On the other hand, the use of classical statistical methods is complicated by the sparseness of data and the lack of a standard annotated disambiguated corpus. Thus, the use of each of these methods is not sufficiently effective.

One possible solution to this problem is described in [7]. The method was used for disambiguation of texts on the Turkish language, where the number of wordforms with multiple parsing options, reaches 40%. According to the results of this work, the accuracy of the method for the Turkish language reached 96% (with an accuracy of classical statistical methods of 91%). Typological and genetic proximity of the Turkish and the Tatar language suggests that this method is able to show good results for the Tatar language.

As well as in the Tatar language, in the Turkish language the number of possible types of homonymy is not limited, which in turn leads to failure when using classical statistical methods due to the sparseness of data. To avoid this, instead of searching for the contextual constraints for each type of homoforms, the algorithm searches for contextual constraints for each morpheme, the

number of which is limited, in contrast to the number of types of homoforms: 126 morphemes for the Turkish language [1] and 120 morphemes for the Tatar language [4]. It is obvious that this approach significantly reduces data sparseness.

According to this method, training data is collected for each morpheme from the sample of wordforms, which contain the given morpheme at least in one of the possible morphological parses. The received data are classified as "positive" or "negative", depending on whether the morpheme is included into the contextually suitable paradigm. On the basis of these data and using a special algorithm, the grammatical disambiguation rules are trained [1].

In order to predict a suitable parsing option of an unfamiliar wordform, the morphological analyzer firstly analyzes the wordforms to the greatest possible extend by all possible paradigms. Next, on the basis of rules, for each morpheme a certain probability of its presence or absence in the given wordform in the given context is defined. The final result is calculated taking into account the accuracy of each rule, and ultimately the most likely parse is selected [1]. A distinguishing characteristic of this model and the learning algorithm (GPA algorithm) is their high resistance to irrelevant and redundant features.

The problem of the lack of a fully annotated disambiguated corpus of the Tatar language, which would be used as training data, can be partially solved by choosing for analysis not the homoforms with a certain morpheme, but on the contrary, the wordforms with the given morpheme and a single parsing option. This will allow to identify the contextual constraints directly for the morpheme. However, this approach does not cover the entire set of morphemes (e.g., the morphemes, for which there have not been found worforms with a single parsing option). In such cases, contextual rules are designed manually or after a complete annotation of the model fragment of the corpus.

6 Software Modules for Context Rules Developement

As part of this research, we have developed a software tool designed to create, edit and test the database of context rules for the tasks of automatic grammatical disambiguation in the Tatar language [8].

This module can be used both separately (for this, contextual disambiguation rules should be designed for all types of homonyms), and in combination with the probabilistic and statistical methods. The second part of the toolkit "LangRuleBase-PMM module" [8] uses this database of context rules for grammatical disambiguation in texts. This kind of toolkit, which takes into account the particularities of the Tatar language, was developed fort he first time. It is aimed at assisting the research work of a philologist.

To facilitate the annotation process of the Tatar language corpus (including manual disambiguation), as well as to provide convenient access to the statistical data of the corpus, we developed a web application that makes the work with corpus texts more convenient and flexible for statistical research. This software module, in addition to the possibility of expanding the corpus and morphological annotation, supports the option of manual grammatical disambiguation.

7 Conclusion

Formal context-oriented classification of homoforms and development of context rules for grammatical disambiguation using experimental statistical data in the Tatar language have been carried out for the first time. Linguistic resources and software modules developed on the basis of the classification and context rules allow to

perform disambiguation in the Tatar National Corpus and other applications. Estimated cumulative effect in the case of disambiguation of the identified frequent types of homonymy in the Tatar language corpus can be up to 50%. Our future research will be focused, on the one hand, on the study of disambiguating contexts and the development of contextual disambiguation rules and, on the other hand, on the analysis of statistical regularities in the field of polysemy at different language levels and the search for effective approaches to disambiguation taking into account the particular characteristics of the Tatar language.

8 Acknowledgements

The work is supported by the Russian Foundation of Basic Research and the Government of the Republic of Tatarstan, (project # 12-07-97015)

9 References

[1] Yuret D., Ture F. *Learning Morphological Disambiguation Rules for Turkish*. Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. NewYork, 2006. Pp. 328–334.

[2] Galieva A.M., Khakimov B.E., Gatiatullin A.R. *Metayazyk opisaniya struktury tatarskoy slovoformy dlya korpusnoy grammaticheskoy annotatsii* [Metalanguage of description of a Tatar wordform for corpus-based grammatical annotation]. Proceedings of the Kazan University, Ser. Human sciences, 2013. V. 155, b. 5. Pp. 287-296. (rus)

[3] Nevzorova O.A., Zinkina Yu.V., Pyatkin N.V. *Razresheniye funktsionalnoy omonimii v russkom yazyke na osnove kontekstnykh pravil* [Resolution of functional homonymy in the Russian language based on context rules]. Proceedings of "Dialog'2005" International Conference. Moscow: Nauka, 2005. Pp. 198-202. (rus)

[4] Suleymanov D.Sh., Gilmullin R.A. *Dvukhurovnevoye opisaniye morfologii tatarskogo yazyka* [Two-level description of the Tatar language morphology]. Proceedings of "Language semantics and image of the world" International Scientific Conference. Kazan: Ed. Kazan State University, 1997. Vol 2. Pp. 65-67. (rus)

[5] Suleymanov D.Sh., Gilmullin R.A., Gataullin R.R. *Programmnyy instrumentariy dlya razresheniya morfologicheskoy mnogoznachnosti v tatarskom yazyke* [Software toolkit for morphologic disambiguation in the Tatar language]. Proceedings of OSTIS-2014 IV International scientific and technical conference. Minsk, 2014. Pp. 503-508. (rus)

[6] Suleymanov D.Sh., Khakimov B.E., Gilmullin R.A. *Korpus tatarskogo yazyka: kontseptualnyye i lingvisticheskiye aspekty* [Tatar language corpus: conceptual and linguistic aspects]. Bulletin of Tatar State Humanitarian Pedagogical University. 2011. № 4 (26). Pp.211-216. (rus)

[7] "Tugan Tel" Tatar National Corpus. – URL: http://web-corpora.net/TatarCorpus/search/?interface_language=ru.

[8] Khakimov B.E., Gilmullin R.A. *K razrabotke morfologicheskogo standarta dlya sistem avtomaticheskoy obrabotki tekstov na tatarskom yazyke* [Notes on the development of a morphological standard for automatic text processing systems in the Tatar language]. System analysis and semiotic modeling: Proceedings of all-Russia conference with international participation (SASM-2011). Kazan, 2011. PP. C. 209-214. (rus)