EPiC Series in Language and Linguistics

Volume 1, 2016, Pages 228–235

CILC2016. 8th International
Conference on Corpus Linguistics

EPiC
Language
and Linguistics

Grammatical Disambiguation in the Tatar National Corpus

Bulat Khakimov[1,2], Ramil Gataullin[2,1] and Rinat Gilmullin[2,1]
[1] Kazan Federal University, Russia
[2] Applied Semiotics Research Institute of the Tatarstan Academy of Sciences
bulat.khakeem@gmail.com, ramil.gata@gmail.com,
rinatgilmullin@gmail.com

Abstract

This paper concerns the issues of grammatical ambiguity in the Tatar National Corpus and the possiblities for automation of the disambiguation process in the corpus. Grammatical ambiguity is widely represented in agglutinative languages like Turkic or Finno-Ugric. In order to build the grammatically disambiguated subcorpus, we have developed a special software module which searches for ambiguous tokens in the corpus, collects statistical information and allows creating and implementing the formal disambiguation rules for different ambiguity types. Disambiguation in the corpus is based on the context-oriented classification of ambiguity types which has been carried out on statistical corpus data in the Tatar language for the first time. We can say that we use the corpus as a source of our research and at the same time as a destination for implementing the results. Estimated cumulative effect of disambiguation of the identified frequent ambiguity types in the Tatar National Corpus can be up to 50%.

1 Introduction

The problem of grammatical ambiguity and its resolution is one of the most pressing problems in modern computer and corpus linguistics (Yuret & Ture, 2006). “Tugan Tel” Tatar National Corpus was developed in the “Applied semiotics” Research Institute of the Tatarstan Academy of Sciences and the Kazan Federal University (Suleymanov et al, 2013). It employs the system of automatic morphological annotation on the basis of our own morphological analyzer (Suleymanov & Gilmullin, 1997). In order to adequately reflect the specifics of the Tatar language, a morphological standard of the corpus was developed (Khakimov & Gilmullin, 2011). Research on specification and improvement of the metalanguage for the grammatical corpus annotation of Tatar wordforms is currently carried out (Galieva et al, 2013). The general conception of the corpus is presented in (Suleymanov et al, 2011). To implement the grammatical disambiguation in the Tatar National Corpus, developers have conducted a study of contextual constraints of different types of grammatical

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 228–235

homonyms, involving statistical corpus data, and suggest the methods of automatic grammatical disambiguation for the Tatar language.

The Tatar language belongs to the Turkic group that forms a subfamily of Altaic languages. The Tatar language is spoken in West-central Russia (in the Volga region) and southern parts of Siberia. The number of Tatars in Russia in 2010 was 5.31 million people.

The Tatar National Corpus has a system of grammatical annotation that is oriented at presenting all the existing grammatical word-forms. Grammatical annotation of a Tatar word includes the information about the part of speech of the word and a set of morphological features. Morphological annotation is carried out using our own morphological analyzing tool which was created on the basis of the PC-KIMMO two-level morphology model. The search functionality of the Corpus includes search queries for lemmas (lexemes), word forms, and individual grammatical features.

2 Statistical Information about the Corpus Data

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 limited subcorpus of the Tatar National Corpus. The total volume of the subcorpus was 21,940,452 tokens, and the proportion of tokens with alternative parses was 25.75%.

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

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 Homonymy Types Relevance Evaluation

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 excluding 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 Identifying Frequent Types of Grammatical Homonymy on the Corpus Data

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 Methodology of Disambiguation

As an analysis of the existing methods shows, the problem of morphological disambiguation has been solved by researchers in different ways. The first algorithms were based on the rules. Later statistical algorithms started to be applied. One of such methods based on a Markov model is already considered to be classic. Many methods are language independent. However, each language has its own characteristics, taking into account that one can achieve more exact results.

Belonging of the Tatar language to agglutinative type determines the characteristics of morphological structure of the word: word forms are formed by adding word-formation and inflection affixes to the stem. Each grammatical meaning is usually expressed by an individual affix. Taking into account these features, for the Tatar language, it has been developed an adapted morphological analyzer on the basis of two-level model of the morphology of the Tatar language.

According to statistic data from the corpus of the Tatar language, the most important in morphological disambiguation in the Tatar language is the problem of resolving functional homonymy.

As the research results show, for the model of morphology of the Tatar language we can expect effective applicability of both contextual and statistical methods (Khakimov et al, 2014). The software tools developed on the basis of contextual rules proved to be efficient and showed acceptable results for the Tatar language (Gataullin & Gilmullin, 2015).

The academic book of Tatar grammar (1992, vol. 3) represents the main formal grammatical models of Tatar word-combinations (15 basic, 80 particular types) indicating the main and dependent words, means of linking, grammar, and in some cases, lexical semantics. These models can serve as a basis for determining the resolving contexts. A certain strictness of the agglutinative syntactic structure allows expecting the detection of clear context restrictions.

However, a rule-based approach, as our previous studies (Khakimov et al, 2014; Gataullin & Gilmullin, 2015) have shown, is an extremely time-consuming, requires a thorough linguistic examination of each type of homonyms. Absolutely exact resolution of homonyms on the basis of context rules does not seem to be possible for many reasons.

One of them is a complicated structure of the rules of the method of context resolution. We apply a method of contextual disambiguation of functional homonymy which involves several stages:

1) Full classification of the functional homonyms;

2) The selection of a minimal set of resolving contexts for each type. Minimal set means that for each type of functional homonym, the complexity of recognition of each part of speech that belong to this type should be evaluated. Then it is needed to build a set of resolving contexts (SRC) with minimum complexity of recognition. In algorithmic writing, this requirement is expressed by the following rule: if to a functional homonym X having type T1 or T2, a rule from SRC is applicable, the homonym type X is determined by the applied rule, or an alternative type is attributed;

3) Building a management structure of a generalized rule ensuring maximum recognition accuracy.

For each type of functional homonymy, it is developed generalized rule for disambiguation of this type of homonyms. The generalized rule is an ordered set of rules, written in a special formal language. Each rule within a set fixes some resolving context. The structure specifies the order of rules, which is based on an evaluation of frequency of the contexts.

Let's consider the resolution of functional homonymy of the following type:

(V + Refl) / (N + 3 PossSg + Acc), where

(V + Refl) is a verb with affix of reflexive and passive voice, and

(N + 3PossSg + Acc) is a noun with affixes of the 3d person singular.

Example of the homonym: '*asılın*'.

The variants of affixal structure of the homoform:

(1) as(V)+Il(Refl)+In(Refl) '*hang*'

Like in: *Muyınğa asılın* 'Hang on the neck')

(2) asıl(N)+SI(3PossSg)+nI(Acc) 'the essence of something'

Like in: *Ğömerneň asılın aňlağız* 'Understand the essence of life'

The potential models, the main components and semantics of word combinations:

(1) not found as a subordinate component

(2) N+Acc→V (main component is a verb, semantics of direct object).

This type of affixal homonymy is resolved as follows:

'If there is a verb in the right context, the potential model of word combination N+Acc→V (2) is possible.

Correspondingly, if this model of word combination is realized, homonymy is resolved according to the 2-d variant of morphemic structure, that is, N+3PossSg+Acc: asıl(N)+SI(3PossSg)+nI(Acc).

The method of functional homonymy resolution based on the context rules uses syntactic patterns. This fact determines the advantages and disadvantages of the method. Primarily, a higher accuracy compared with probabilistic methods should be referred to the advantages. Drawbacks also follow from the nature of its syntax. In practical applications, oriented to concrete sublanguages, it is advisable to use an engineering approach with its pragmatic solutions - isolation of a group of potential homonyms, i. e. almost not homonymous; accounting of semantics of a concrete subject area, actually removing most of the ambiguity, etc. However, it should be realized that such solutions do not reduce the complexity of the phenomenon, but only partially decrease the complexity of recognition accuracy.

The main disadvantage of the method of context rules is the complexity in defining context rules (or context restrictions) for each type of homonymy. Despite the fact that the theme of homonyms in the Tatar language has been well studied (Kurbatov, 1959; Salimgarayeva, 1971; Salakhova, 2007), finding the minimum context restrictions proves to be not obvious task. The more so because the classification of homonyms can always be developed further deepening and detailing the context.

Application of statistical and probabilistic methods for the Tatar language is complicated only by the lack of the tagged corpus for learning. Currently, such subcorpus is being compiled. But proceeding from the fact that for many languages (and Tatar is unlikely to be exception), the statistical methods are inferior in accuracy of the method, based on the rules, it has been decided to combine these methods so that to apply the developed rules for the most simple and obvious cases of ambiguity and for the others - non-obvious – the statistical and probabilistic method. We can judge about applicability and effectiveness of this approach unambiguously only after experimenting with the tagged corpus.

The need for high accuracy (not less than 95%) at a resolution of morphological ambiguity is conditioned by the necessity of the results of this analysis in lexical disambiguation. The task of lexical ambiguity resolution is the next step in the development of computational linguistics for the

Tatar language.

6 Automatic Development of the Grammatical Disambiguation Rules

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 subcorpus, which includes more than 21 million tokens, there are more than 7000 types of homonyms. 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 (Yuret & Ture, 2006). 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 (Yuret & Ture, 2006) and 120 morphemes for the Tatar language (Khakimov & Gilmullin, 2009). 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 (Yuret & Ture, 2006).

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. 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 identifying 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.

7 Software Tool for Creating Context Rules of Grammatical Disambiguation

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 (Gataullin & Gilmullin, 2015).

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 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.

8 Conclusion

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

References

Yuret, D., Ture, F (2006). *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.

Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., Khakimov, B. (2013). *National Corpus of the Tatar Language “Tugan Tel”: Grammatical Annotation and Implementation*. In: Procedia – Social and Behavioral Sciences. Vol.95 (2013). pp. 68-74.

Nevzorova, O., Zinkina, Y., Pyatkin, N (2005). *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 (In Russian).

Tatar Grammar (1993). In 3 volumes. Kazan: Tatar publishing company, 1993. V. 2: Morphology. 397 p. (In Russian).

Tatar Grammar (2002). In 3 volumes. Moscow: Insan, Kazan: Fiker, 2002. V. 2. – 448 p. (In Tatar).

Galieva, A., Khakimov, B., Gatiatullin, A. (2013). *A Metalanguage for Describing theStructure of Tatar Word Forms for Corpus Grammatical Annotations.* Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki, 2013, vol. 155, no. 5, pp. 287-296. (In Russian).

Suleymanov, D., Gilmullin, R. (1997). *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 (In Russian).

Suleymanov, D., Gilmullin, R., Gataullin, R. (2014). *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 (In Russian).

Suleymanov, D., Khakimov, B., Gilmullin, R. (2011). *Korpus tatarskogo yazyka: kontseptualnyye i lingvisticheskiye aspekty* [Tatar language corpus: conceptual and linguistic aspects]. Philology and Culture. 2011. № 4 (26). pp.211-216 (In Russian).

Khakimov, B., Gilmullin, R. (2011). *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 (In Russian).

Gataullin, R., Gilmullin R. (2015). *Web-Interface for Removing Morphological Ambiguity in the Corpus of the Tatar Language.* Open Semantic Technologies for Intelligent Systems OSTIS-2015 Proceedings of IV International Scientific and Technical Conference (Minsk, February 19-21, 2015). – Minsk: BSUIR, 2015, pp. 451-454 (In Russian).

Kurbatov, K. (1959). *Grammatical Homonyms in the Tatar Language*. In: Journal of Tatar Language and Literature. – Kazan, 1959, pp. 307-311 (In Tatar).

Salakhova, R. (2007). *Homonym Suffixes of the Tatar Language*. – Kazan: Gumanitarya, 2007, 204 p. (In Russian).

Salimgarayeva, B. (1971). *Homonyms in Modern Tatar Language: Abstract of Dissertation*. – Ufa, 1971, 82 p. (In Tatar).

Suleymanov, D. (1994). *Regularity of Morphology of the Tatar Language and Types of Violations in the Language*. Cognitive and Computational Linguistics / Eds: R.G. Bukharayev, V.D. Solovyev, D. Sh. Suleymanov. Kazan: KSU, 1994, pp. 77-106 (In Russian).

Weischedel et. al (1993). *Coping with ambiguity and unknown words through probabilistic models.* Compututational Linguistics. Cambridge, MA, USA: MIT Press, Volume 19 Issue 2, June 1993, pp. 361–382.

Khakimov, B., Gilmullin, R. Gataullin, R. (2014). *Grammatical Disambiguation in the Corpus of the Tatar Language.* Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki, 2014, vol. 156, no. 5, pp. 236-244. (In Russian).

Tatar National Corpus. http://corpus.antat.ru.