Hybrid approach to zero pronoun resolution in Bulgarian

Diana Grigorova

Abstract: *In this paper we present a hybrid approach to zero pronoun resolution in Bulgarian. The pre-processing of text is accomplished by a parser, based on dependency grammar. Machine learning techniques are applied to predict the type of the clause: with a zero pronoun, with a subject or impersonal. Finally, we apply rules using the full syntactic tree provided by the parser to identify the most probable antecedent of the zero anaphor.*

Key words: *anaphora resolution, zero pronoun resolution, machine learning, dependency grammar.*

INTRODUCTION

The anaphora resolution is an issue challenging the computational linguistics specialists since the time of the first developments related to the computer processing of natural language texts. The interest towards this issue stems from the fact that the anaphora, as a phenomenon, is intrinsic to every natural language. It contributes to the proper expression of sequences, dependencies and logical links intrinsic to the human thinking process.The quality computational processing of natural language texts implies recognizing this phenomenon and resolving the ambiguities that go with it. The anaphora resolution is of significant importance in areas like machine translation, automatic summarization, question answering, test generation, etc.

No universal algorithms exist that could determine if two or more phrases are related to one and the same entity. The anaphora resolution is part of the computational linguistics aiming to solve this problem. The word or phrase which points back is called anaphor and the entity to which it refers is its antecedent. The process of determining the antecedent of the anaphor is called anaphora resolution [15]. Among the various types of anaphoric expressions, the most researched one is the pronominal anaphora. In some languages, including Bulgarian, there exist linguistic situations where the coherence of a sentence benefits from the absence of linguistic forms. This class of anaphora is called zero anaphora. "Zero pronominal anaphora occurs when the anaphoric pronoun is omitted but is nevertheless understood" [15]. Below is an example of a sentence with a zero pronoun in Bulgarian:

Маймуните изядоха бананите, защото **[zp]** бяха гладни.

in English: The monkeys ate the bananas because **they** were hungry

The sentence is complex, consisting of two simple sentences (clauses), separated by comma. The symbol "zp" stands for the missing pronoun (zero pronoun). Although the second clause is subordinated, the subject there is omitted, and nevertheless this sentence is absolutely acceptable in Bulgarian. The pronoun "те" (they), which would point back to the word „маймуните" (monkeys) is missing, but it is fully implied. If the sentence is translated into English word by word, without resolving of the zero anaphor, the result would be:

"The monkeys ate the bananas because were hungry."

This sentence is definitely not correct in English. The absence of subject in the second clause is unacceptable.

EXISTING SOLUTIONS

Many of the existing solutions for pronominal anaphora resolution use as starting point the remarkable and important work of Lappin and Leass [13]. Lappin and Leass suggest an algorithm for resolution of the third person personal pronouns. Their algorithm is based on the presumption that a full syntactic tree based on context-free grammar is available. Lately, a lot of effort has been put into the development of robust and reliable algorithms, deliberately working in a limited knowledge environment. Solutions relying on limited knowledge are being tendentiously sought. Partial (shallow) syntactical parsing recognizing only certain phrases (usually noun phrases) in the sentence is often suggested instead of the full syntactical parsing [12, 14].

At the same time, another approach for solving the problem of pronominal anaphora resolution is being developed: machine learning as implied by the corpus linguistics [1, 5]. The approaches based on machine learning are considered attractive, as their use reduces the efforts for developing algorithms identifying the antecedent. But let us not forget that they rely on corpora with significant volume, which have been previously annotated by specialists and processed.

Although most of the existing algorithms have been developed for the English language [1, 5, 12, 13], there are algorithms developed for other languages as well. One of the most significant and quoted works for Spanish language is [16]. It presents a rule-based system for identifying the antecedent of personal, demonstrative, reflexive and omitted pronouns (zero pronouns).

There is a similar text processing system for Bulgarian, which includes a module for anaphora resolution [20]. This module searches the antecedent of personal pronouns in third person and represents an adaptation of the method suggested by Mitkov, which is in the category of the intentionally knowledge-poor approaches [14]. The zero anaphor is not discussed in this work.

The zero pronouns are of special interest for the languages they are intrinsic to: Spanish, Romanian, Portuguese, Chinese, Korean, Bulgarian, etc. The zero anaphor resolution is always achieved in two steps: identifying the sentences with zero pronouns and finding the antecedent of the missing anaphor. The proliferation of zero pronouns in Romanian is discussed in [3], and a rule-based approach for identifying sentences with zero anaphor in Spanish – in [17]. Finding a simple sentence (clause) with a zero pronoun is a classification task, which can be solved using machine learning [4, 18]. The hybrid approach – defined as combination of rules and machine learning – is an attractive idea used for pronominal anaphor resolution (but not for zero anaphor resolution) in [2, 9, 19]. Isozaki and Hirao in [10] use ordered rules in combination with machine learning for identifying the antecedent of a zero anaphor in Japanese, in case the positions of the zero pronouns have already been found.

The zero-pronominal anaphor in Bulgarian is researched in [6]. Heuristic algorithms, based on rules for tagging the sentences with zero pronouns and finding their antecedent, are suggested. A parser based on context-free grammar is used for that purpose.

The related research proved that the process of creating the rules is an extremely difficult, time-consuming and error-prone task. This is especially true for the rules for identifying clauses with zero anaphor. Machine learning techniques have been used in order to make the solution of this task more reliable and resilient to exceptions [7]. The concept to be learned is the type of clause. This is a classification task which provides three distinct values of the class – clause with a subject, clause with a zero pronoun and impersonal clause.

PARSER

The parsers based on context-free grammar are presupposed by the idea of Chomsky, according to which the structure of the sentence consists of subject and predicate. The subject is expressed by a noun phrase and the predicate by a verb phrase. Such grammars are called phrasal or constituent grammars [11]. Their main feature is the idea of constituency – each phrase may consist of other phrases. The main disadvantage of the parser, based on constituent grammar is the explosion of grammar rules and a corresponding loss of generality, which occur when the diversity of natural language have to be put into context-free grammar rules.

In order to create a large training file containing more than 7000 instances of clauses, we used a parser, provided by the colleagues from the Linguistic Modeling Department of the Bulgarian Academy of Science. This parser is based on dependency grammar. This type of grammar is especially suitable for languages with free word order, as is the case with Bulgarian. The dependency grammar usually assigns the main function in the sentence to the verb and accepts it as a root. It does not depend on any other word in the sentence, while they are related to it through dependencies. The parser generates a full syntactic tree and provides very rich linguistic information about the words in the sentence and about the dependencies between them. The dependency relations are shown in Table 1.

Name	Description
punct	Punctuation
clitic	Clitic form
mod	Modifier (dependants which modify nouns, adjectives, adverbs)
prepcomp	Complement of preposition
comp	Complement (arguments of: non-verbal heads, non-finite verbal heads,
adjunct	Adjunct (optional verbal argument)
subj	Subject
xadjunct	Clausal adjunct
xsubj	Clausal subject
xmod	Clausal modifier
xcomp	Clausal complement
xprepcomp	Clausal complement
conj	Conjunction in coordination
conjarg	Argument (second, third,…) of coordination
pragadjunct	Pragmatic adjunct
marked	Marked (clauses, introduced by a subordinator)
obj	Object (direct object of a non-auxiliary verbal head)
indobj	Indirect object (indirect argument of a non-auxiliary verbal head)

Table 1

CLASSIFIER

As we mentioned before, zero pronoun resolution is a two-step task. First, the clause with zero anaphora has to be marked, and second, the antecedent of zero anaphor has to be found.

Machine learning technique is used for the first step task. It is described in detail in [7]. The experiments were carried out with the Weka package [21]. Here we will note the most important facts and outcomes.

The training file consists of 7009 instances. Each instance is a verb in a clause. The classifier has three possible outputs: clause with a subject, clause with a zero pronoun or impersonal clause. 51.8% of the verbs have subject, 38.8% have zero pronoun, and for 9.4% of the sentences the subject is impossible, i.e. the sentence is impersonal. This distribution matches the distribution of the respective types of sentences in Bulgarian [8]. The texts in training file are retrieved from the web and from digitalized book, encompassing the literary and news genres. It is important to note that the classifier has not been trained with randomly chosen continuous texts, but with selected chunks of text and sentences from the three types in the mentioned ratio.

The feature vector consists of 16 characteristics. The first 15 have been extracted from the XML file generated by the parser, while the last one has been entered manually by the annotator. The annotator enters one of the three possible values of the class.

The JRip algorithm produces the best results. This is an algorithm which extracts rules. The algorithm has been tested through a ten-fold cross-validation. The achieved accuracy (correctly classified instances) is 92.71%. Table 2 presents the precision, recall and F-measure values.

Clause	Precision	Recall	F-measure
Subject	0.953	0.952	0.953
With ZP	0.919	0.919	0.919
Impersonal	0.821	0.822	0.821

Table 2

IDENTIFYING THE ANTECEDENT

Our goal now is to find clauses with zero pronouns using the results from machine learning and to determine the antecedent of zero anaphor by applying rules. The rules exploit the parsing trees of the current sentence. We use raw texts from Bulgarian authors, parsed by the parser and presented in XML format.

The first approach to leverage machine learning is to use the chosen raw text as application (test) set in the machine learning process. The idea is attractive, but it makes Weka package immutable part of the zero anaphora resolution system. This idea also presupposes the parsed text to be limited in size.

The second approach to leverage machine learning is to take the rules, extracted from JRip algorithm, shown in Fig.1, and apply them directly to the parsed text. We chose this way to process the text for marking the clauses with zero pronouns. When a clause with zero pronoun has been found, rule-based algorithm determines the antecedent. The execution flow is shown in Fig. 2.

In order to find the antecedent, we need to know its most important features. Zero pronoun distribution in Bulgarian and its features have been presented in [8]. The zero anaphor in Bulgarian can only be in a subject position, but the position of the antecedent can vary. We are interested in the syntactic position of the antecedent: subject, direct object, indirect object, uncoordinated attribute, adjunct phrase. The analysis of a corpus consisting of texts from different genres, exceeding 35500 words and 1000 sentences with zero pronoun, proved that in 91,59% of the cases the antecedent of the zero anaphor is the subject in one of the preceding clauses [8]. Parsers, based on CFG do not provide information about the syntactic role of the words. Such parsers produce a syntactic tree, composed by phrases: verb phrase, noun phrase, preposition phrase, etc. The most

important question in this context is: which noun of the preceding noun phrases is the antecedent, i.e. which noun is the subject? In order to find the subject, we first need to pass the potential candidates to person, number, and if necessary, gender agreement test. If there is more than one candidate for antecedent, some preferences are applied. A strategy, known as a list of preferences [16, 6] can be used.

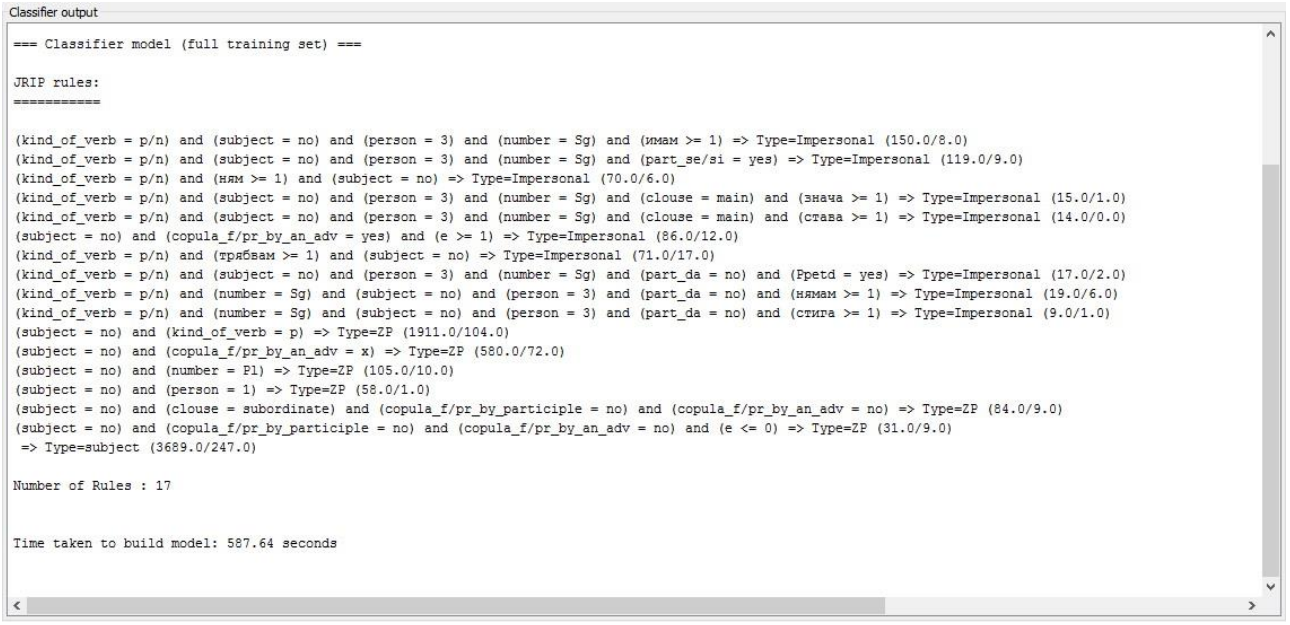

Fig.1 JRip rules

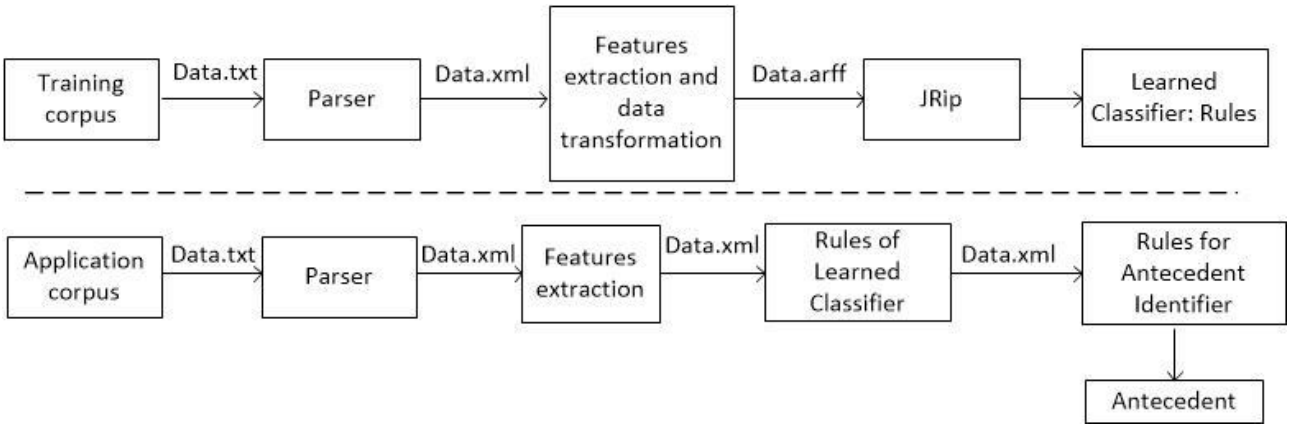

Fig.2 Execution flow

The main advantage of the dependency-based grammar parser is its capability to assign the "subject" attribute. Our algorithm traces links of the verbs, marked as having zero pronoun. If these links traversing the words in the sentence, point to a verb, which is joined to a word, marked as subject, this word is specified as antecedent. The algorithm points to a single word and therefore we do not need any rules for ranking more than one candidate.

EVALUATION

In [15] Mitkov distinguishes evaluation of anaphora resolution algorithms and evaluation of anaphora resolution system. Our purpose is to evaluate the overall performance of the system. The system includes parser, machine learning and rules for finding the antecedent. The main sources of errors for the whole systems are the parser, the feature vector, the annotator, the training data, machine learning, the rules and the evaluation data. Another source of errors is the fact that we search for the antecedent only in the current sentence. In our evaluation texts the percentage of antecedents, whose antecedents are in other sentences is about 10% (including cataphora).

Recall and precision have been adopted by a number of researchers for evaluation of anaphora resolution algorithms or systems. Baldwin (1997) defines recall and precision in the following way:

$$Recall = \frac{Number\ of\ correctly\ resolved\ anaphors}{Number\ of\ all\ anaphors}$$

$$Precision = \frac{Number\ of\ correctly\ resolved\ anaphors}{Number\ of\ anaphors\ attemped\ to\ be\ resolved}$$

We should note that in our evaluation the "number of all anaphors" and the "number of correctly resolved anaphors" have been identified manually.

The evaluation has been performed on a corpus of 50 pages of 2 novels from Bulgarian authors. The text is raw, without any pre- or post-editing. It consists of 25975 words and 475 zero pronouns.

Precision is 72,6%. The rules give 100% precision when they are applied on correctly marked clause, whose antecedent is in the current sentence. But we detected the following sources of errors:

1) Certain cases of zero pronouns in singular or plural with exophoric antecedent. The antecedent is exophoric when it exists in the real world, but is not lexically presented. Some of the exophoric antecedents are correctly identified – when there is no potential subject in the sentence. But in some cases the parser corelates the verb having zero pronoun, whose antecedent is exophoric, with a verb which has a subject. This subject is pointed to be the antecedent and this is a mistake;
2) The cases, when the antecedent is not in the current, but in other sentence. Direct speech is a common origin of sentences with inter-sentential zero anaphora;
3) Zero pronoun in plural, whose antecedent is a group of coordinated antecedents in singular or collective noun.

Recall is 46,8%. Sources of errors, intrinsic to automatic resolution are added in calculation of the recall. These are:

4) The parser. In some cases the parser does not correctly specifies the subject, especially when the subject follows the predicate or when the predicate is in passive voice.
5) The machine learning. A pre-requisite for the supervised machine learning is the existence of manually annotated set of data. Errors in manual annotation and the choice of the characteristic, constituting the feature vector influence the quality of the machine learning. The rules, extracted by the JRip algorithm do

not always correctly specify clauses with zero pronouns. The upper limit of the accuracy of the system is the upper limit of the accuracy of the training data.

CONCLUSIONS AND FUTURE WORK

The obtained results can be regarded to be satisfactory and promising. As a first step of improvement we will formulate some rules for identifying of coordinated arguments of the subject. Collective nouns constitute a restricted set and they can be put in the list, which will be tested during the process of resolution of 3rd person plural zero pronoun. The expectations are for improvement of precision.

Future efforts could also be focused on using variations of training data sets and eventually on certain changes in the feature vector for different genres. Evaluation tests on different genres could be carried out as well.

REFERENCES

[1] Aone C., and Bennet S., 1996, Applying machine learning to anaphora resolution", Connectionist, statistical and symbolic approaches to learning for Natural Language Processing, p. 302-314, Berlin, Springer

[2] Barbu, C., "Bilingual pronoun resolution: Experiments in English and French", PhD Thesis, University of Wolverhampton, 2003

[3] Claudiu, M., Ilisei, I and Inkpen, D., "Romanian Zero Pronoun Distribution: A Comparative Study" http://clg.wlv.ac.uk/papers/Mihaila-Ilisei-Inkpen-LREC2010.pdf

[4] Claudiu, M., Ilisei, I and Inkpen, D., "To be or not to be a zero pronoun: A machine learning approach for Romanian" http://clg.wlv.ac.uk/papers/ilisei-PROMISE-10.pdf

[5] Connolly D., Burger J. and Day D., 1994, "A machine learning approach to anaphoric reference", Proceedings of the international conference on new methods in language processing (NEMLAP)

[6] Grigorova, D. „Heuristic algorithm and machine learning approach to zero pronoun resolution in Bulgarian", PhD thesis in Bulgarian, TU Sofia, 2014

[7] Grigorova D., "A machine learning approach for identifying zero pronouns in Bulgarian", Proceedings of the 15th International Conference CompSysTech'14, Ruse Bulgaria June 27 2014

[8] Grigorova, D., "Zero pronoun distribution in Bulgarian: a comparative study", Information Technologies and Control, №1/2012, p.29-36

[9] Hinrichs, E. W., Filippova, K., and Wunsch, H. "A Data-driven Approach to Pronominal Anaphora Resolution in German". Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing (RANLP 2005), p. 239–245.

[10] Isozaki, H. and Hirao., T., "Japanese Zero Pronoun Resolution based on Ranking Rules and machine Learning", Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03), p.184-191.

[11] Jurafsky, D. and Martin J.H., Speech and language processing, Prentice Hall, 2000

[12] Kennedy, C and Boguraev, B. "Anaphora for everyone: pronominal anaphora resolution without a parser", Proceedings of the 16th international conference on computational linguistics 1996, p. 113-118, Copenhagen, Denmark

[13] Lappin, S. and Leass, H., "An algorithm for pronominal anaphora resolution" Computational Linguistics, 1994 20(4) p. 535-561

[14] Mitkov R., "Robust pronoun resolution with limited knowledge", Proceedings of the 17th International Conference on Computational Linguistics, 1988, p.869-875, Montreal, Canada

[15] Mitkov R., Anaphora Resolution, Pearson Education, 2002

[16] Palomar M., A. Ferrandez, L. Moreno P. Martinez-Barco, J. Peral, M. Saiz-Noeda, R. Munoz, "An Algorithm for Anaphora Resolution in Spanish Texts", Computational Linguistics Volume 27, Number 4 pp. 545-567, 2001

[17] Rello L., and I. Ilisei, "A Rule-Based Approach to the Identification of Spanish Zero Pronouns" Student Research Workshop, RANLP 2009 – Borovetz, Bulgaria, pages 60-65

[18] Rello L., "Elliphant: A machine learning method for identifying subject ellipses and impersonal constructions in Spanish", A thesis submitted for the degree of Erasmus Mundus International Master in Natural Language Processing and Human Language Technilogy, 2010.

[19] Stuckardt, R., "Machine-learning-based vs. manually designed approaches to anaphor resolution: the best of two worlds", Proceedings of 4th Discourse anaphora and anaphor resolution colloquium, University of Lisbon, Sept. 2002, p. 211-216

[20] Tanev, Hr., and Mitkov, R., "Shallow language processing architecture for Bulgarian", Proceedings of the 19th International conference on Computational linguistics - Volume 1 Taipei, Taiwan, pp. 1-7, 2002

[21] http://www.cs.waikato.ac.nz/ml/weka/

ABOUT THE AUTHOR *

Diana Grigorova, Assoc. prof., Ph. D., Technical University of Sofia, E-mail: dgrigorova@tu-sofia.bg