An Algorithm for Zero Pronoun Resolution in Bulgarian

Diana Grigorova

Abstract: *This paper presents an algorithm for identifying the noun phrase antecedents of zero pronouns in Bulgarian. The algorithm applies to syntactic representation in XML format, generated by context-free grammar parser. Like the parser, the algorithm is implemented in Java. The author has tested it on Z-corpora, specially created for the investigation of zero pronouns in Bulgarian. The corpora are manually annotated and contain 1029 zero pronouns occurrences. The algorithm has 84% critical success rate. The algorithm is compared to another approach, which has been proposed in literature. The author's algorithm achieves 3% higher critical success rate.*

Key words: *Pronoun resolution, Zero pronoun resolution, Zero pronouns and zero subjects, Z corpora.*

I. INTRODUCTION

Anaphora resolution is of significant importance to a class of applications in natural language processing area, such as machine translation, automatic summarization, question answering and generation of multiple choice tests. There exist no universal algorithms to determine whether two or more phrases relate to the same entity. Anaphora resolution is a field research in Computational linguistics which attempts to resolve the problem. Halliday and Hasan define the term anaphora as the cohesion which points back to some previous item [6]. "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" [10]. The most widespread type of anaphora is that of pronominal anaphora. The pronoun is part of the speech which can replace a noun, an adjective, a numeral or it can be used to point to some objects and their features. The pronouns don't have their own lexical meaning – they realize anaphoric function in the speech, i.e. they direct the thought to already mentioned or to already known objects [18]. The problem, which anaphora resolution algorithm has to solve, is to find the antecedent of the pronoun. But there are many linguistic situations when the coherence of a sentence benefits from the absence of linguistic forms. – Ex. 1. This class of anaphora is called zero anaphora. "Zero pronominal anaphora occurs when the anaphoric pronoun is omitted but is nevertheless understood" [10].

Each example will be presented in two forms: as a Bulgarian sentence and as a translation in English. All examples in the paper are taken from the annotated Z-corpora.

Ex. 1 Като $_{zp}$ *[тя]* стигна до него, *Ирина* отпочина малко, после $_{zp}$ *[тя]* взе тежката кошница в другата ръка и $_{zp}$ *[тя]* продължи към града.

When *she* reached it, *Irina* had some rest, then *she* took the heavy basket in the other hand and *she* continued her way to the town.

The symbol "zp" stands for the missing pronoun (zero pronoun). Ex. 1 illustrates intra-sentential anaphora – the antecedent *Ирина* (*Irina)* is allocated in the main clause and the missing anaphors are in the subordinated clauses of the same sentence. It also gives us an example of cataphora, too – the anaphor in the first clause precedes the antecedent in the second.

Shallow parser, based on context-free grammar, pre-processes the text. Pre-processing includes four steps and covers the current sentence plus the 2 previous. In the first step the sentences are separated; in the second each sentence is further divided into clauses[1] and in the third syntactic analysis is performed. Finally, the result is converted to XML format and passed to the algorithm. The algorithm recognizes the clauses with zero pronouns and finds the antecedents of the missing pronouns. The analysis of created Z-corpora reveals that in more than 80% of the cases, after the test for morphological agreement, there is more than one candidate for antecedent. To choose "the right" antecedent, our first version of the algorithm was realized using the idea of "list of preferences", adopted for Bulgarian [3, 4]. The analysis of the mistakes shows some common features, which we try to overcome with a strategy of assigning scores. Lappin and Leass use such a strategy in [9]. Mitkov uses similar approach in his robust, knowledge poor algorithm [10]. Both algorithms are directed to resolution of antecedents of third person pronouns and do not resolve the antecedents of zero pronouns.

The contribution, presented in this paper is an original algorithm for zero pronoun resolution in Bulgarian, based on the strategy of assigning scores. The algorithm relies on syntactic structures and on a simple mechanism, which registers the changes of attentional state. Neither semantic knowledge is used, nor a model of discourse structure.

This paper is structured as follows: Section II discusses language premises in Bulgarian; section III - the related research on the problem of pronominal anaphora resolution. Section IV describes the Z-corpora. In section V we present the two versions of the algorithm: the first one, based on list of preferences and the second, based on the assignment of scores. Evaluation of the results is presented in section VI and future work – in section VII.

II. LANGUAGE PREMISES

The syntax of Bulgarian language allows clauses, which have only a predicate as main part (they don't have a word in a subject position), to exist as correct and completed.

These clauses can be divided in two types [1, 13, 16]:

1) Clauses, where the inflexion of the verb expresses the subject, although it is not lexically present.
2) Impersonal clauses – subject does not exist.

From the point of view of computational linguistics the clauses from the first type are clauses with zero pronouns. The antecedents can be lexically present, or not. When the antecedent exists in the real world, but it is not lexically present, we classify it as exophoric [2, 14, 15]. Example 2 contains two clauses with zero pronouns with exophoric antecedents.

Ex. 2 - Баща ти заръча $_{zp}$ *[ти]* да се прибереш по видело - предупреди той, когато $_{zp}$ *[те]* се разминаваха.

Your father said *you* should get home before dark – he warned her when *they* were passing each other.

The conclusion, based on the analysis of clauses with zp in Bulgarian from the point of view of the anaphora resolution algorithm is to divide them as such with verb in 1st or 2nd person, singular or plural, and with verb in 3rd person, singular or plural. The clauses from the first type have exophoric antecedent. The clauses from the second type may have lexically present or exophoric antecedent.

The impersonal clause is a clause which does not have a subject. The emphasis is on the action without indication of the doer – Examples 3 a, b, c, d. The impersonal

[1] Clause is a simple sentence as a part of compound, complex or complex-compound sentence

clauses are marked in italic. The verb has constant form in 3rd person singular and it does not express a real subject [1, 13, 16].

Ex. 3 a) Когато наближиха града, на Борис *се стори*, че трябва да я остави.

When they got closer to the town *it seemed* to Boris that he had to leave her.

b) *Но трябва да се запомни*, че хомеопатията не лекува отделните болести като такива, а човекът като цяло

But it has to be remembered that homeopathy does not cure the illnesses as such, but the human as an entity.

c) Все доказателствата са неясни, *абсолютно винаги има нарушения в процедурите*

It is always that the proofs are unclear; *there are always violations in the procedures*

d) Въпросите се натрупаха *и е редно* някой да даде отговор.

The questions piled up and *it is appropriate* that someone should provide the answer.

The impersonal clauses do not have zero pronouns. The algorithm has to distinguish and discard them.

III. RELATED WORKS

Zero pronouns and methods for their resolution are subjects of investigation in some of pro-drop languages – Portuguese, Romanian, Spanish, Japanese, Chinese and Korean. A methodology for creation of Z corpora in Portuguese is discussed in [11]. A comparative study of zero pronoun distribution in Romanian is presented in [2]. There is extensive research for Spanish. The most significant and quoted work is [12]. It presents a rule-based system for antecedent resolution of personal, demonstrative, reflexive and omitted (zero) pronouns. The algorithm is based on constraints and preferences, like the most important proposals for anaphora resolution as those of Lappin and Leass [9], Kennedy and Boguraev [8], Mitkov [10]. In this research, all clauses which do not have noun phrase (NP) before a finite verb are considered to be clauses with zero pronouns. This condition is very relaxed and many impersonal clauses will be identified as false positive examples. The antecedent is sought in the current sentence (before verb phrase) and in the two previous sentences. The requirement for morphological agreement between VP and NP is the constraint. Preferences are heuristic rules extracted from corpora. They are ordered in a decision list and are applied when there is more than one candidate for antecedent after the morphological test. The system is tested on 1029 zero pronouns. In 868 the antecedent is resolved correctly, i.e. 78.9% success rate. A comparative study of Spanish zero pronoun distribution is presented in [15] and in [14] the same authors proposed 9 rules for identifying impersonal clauses.

Japanese, Chinese and Korean are in the category of topic oriented languages. Zero pronouns are widely used in these languages. In Bulgarian, Spanish, Italian, Romanian, zero pronoun can only be in a subject position, but in topic oriented languages besides in subject position, they could be in the position of direct object, indirect object and could fill any nominative arguments of the verb in adjunct phrases. Zero anaphors in Japanese are so widely practiced, that “a model, which first identifies the most likely candidate antecedent for a given zero pronoun and then it judges whether or not the zero pronoun is anaphoric” is proposed in [7]. This model is opposite to the algorithms for other languages, where first the clause is marked as having zero anaphor and then the process for identifying the antecedent starts. Zero pronouns occur also very frequently in Chinese. A study cited in [20] compares the use of overt subjects in English, Chinese, and other languages. It finds that the use of overt subjects in English is over 96%, while this percentage is only 64% for Chinese. Zero pronouns in Chinese, unlike an overt pronoun, provide no such gender or number information. Zero pronouns, which are not explicitly marked in a text, are hard to be identified. Furthermore, even if a gap is a zero pronoun, it may not be co-referential. “All these difficulties make the identification and resolution of anaphoric zero pronouns in Chinese a challenging task” [20]. The authors propose machine learning approach for identification and resolution of Chinese zero pronouns.

Although anaphora resolution has attracted the attention of many researchers and numerous approaches have been developed, we found only one work dealing with this subject for Bulgarian – [19]. This paper presents an anaphora resolver, which is an adaptation for Bulgarian of Mitkov's knowledge-poor pronoun resolution approach [10]. It resolves only third-person personal pronouns. The problem “zero pronoun resolution” in Bulgarian has not been studied there.

IV. Z-CORPORA

Annotated corpora play important role in most of the natural language processing applications. Our first usage of such corpora was to observe patterns and deduce rules for a rule-based anaphora resolver. Further, the same corpora are used to evaluate the two versions of the algorithm and finally they will be used for the implementation of machine learning methods in zero pronoun resolution in Bulgarian.

We had access to the existing annotated corpora described in [17], created in the Linguistic Modelling Department at Bulgarian Academy of Science (BAS). Although the existing corpora are a valuable resource and every word is marked up with detail linguistic information, we took a decision to create our own corpora especially for the purposes of zero pronominal anaphora. The main features which make the existing corpora unsuitable for our goals are discussed in [5].

The corpora consist of full and partial texts retrieved from the web and from digitalized books, encompassing several genres: legal, literary, news and encyclopaedic. The Bulgarian Constitution and the beginning of the Labour Code represent legal text. The literary genre contains texts only from Bulgarian authors: Dimitar Dimov, Dimitar Talev and Zlatko Enev, an author of books for children. The texts in the news genre have been extracted from articles in web newspapers at the end of 2011. Texts with direct speech are avoided. The encyclopaedic genre includes texts from computer, historic and medical literature taken from the web portal BooksBg.org. The corpora contain 1029 zero pronouns, more or less evenly distributed in the mentioned genres. Direct speech is not annotated.

The annotation scheme includes: the omitted pronoun, its antecedent (head noun in the NP), its dependency head (the clause verb on which the ZP depends), the relation (anaphora/cataphora), type of the sentence (simple, compound, complex, complex-compound), type of the clause (main or subordinate), the distance to the antecedent. The zero anaphor in Bulgarian can only be in a subject position, but the position of the antecedent can vary. That is why the syntax position of the antecedent was marked as: subject, direct object, indirect object, uncoordinated attribute, adjunct phrase. This is the final annotation criterion.

The annotation of the corpora was performed by one annotator – the author of this paper. The annotation of the Bulgarian Constitution performed by the author was compared to the annotation of the same text by members of the Linguistic Modelling Department at BAS as part of their corpora. This comparison revealed the differences, which give us the motives for creating new corpora. The remaining cases of ZPs coincide.

Detailed information about the reasons for creating new corpora, annotation criteria and distribution of zero pronouns in Bulgarian can be found in [5].

V. ZERO PRONOUN RESOLUTION

The algorithm begins with marking the clauses with zero pronouns. A feature of clauses of this type is the absence of NP before VP. Clauses, which satisfy this condition, are potentially clauses with zp [2, 12, 14]. In Bulgarian, the noun phrase, which is part of preposition phrase (PP), cannot be a subject of the clause. That is why the clause, which has NP only as part of PP before VP, should be considered as clause with potential zp. Clauses, which satisfy this condition, are divided into 3 groups:

1) Clauses with verb in 1st or 2nd person, singular or plural;
2) Clauses with verb in 3rd person singular;
3) Clauses with verb in 3rd person plural.

The clauses from the first group have exophoric antecedent – the speaker or the hearer. There are three possibilities for the clauses from the second group: clause with zp, impersonal clause or clause with a subject after the predicate. There are two possibilities for the clauses from the third group: clause with zp or clause with a subject after the predicate. Our purpose was to formulate rules to delimit clauses with zp from other types, which are false positives. The most significant criterion is the type of the verb. We distinguish 4 types of verbs: impersonal, personal (finite), auxiliary and verbs with dual use: finite and impersonal. Impersonal constructs in Bulgarian can be expressed by finite verbs. We formulated 12 heuristic rules and 2 exceptions to discard the false positives and not to omit the false negatives. The next example is one of the rules.

Ex. 4 If the verb has dual use and there is no particle "ce" as part of VP, the clause is impersonal.

Exception 1. If the same verb is used in the previous clause as personal, the current clause has zp.

Exception 2. This exception refers to personal and impersonal use of the verb "има" (the Bulgarian version of the verb "to have" has dual use - as finite and as impersonal with the meaning of "there is"). If the verb is „има" and there is no NP before the VP, the clause is impersonal except for the cases:

1) There is a proper name or personal pronoun in the previous clause. In this case the clause is not impersonal, it has zp. The antecedent is the corresponding proper name or personal pronoun.

2) The previous clause has zp and verb in singular. In this case the clause is not impersonal, it has zp and the antecedent is the recovered antecedent from the previous clause.

After the clauses with zp are marked, the second part of the algorithm starts: finding the antecedent. The constraint, applied to the candidates for antecedent is the morphological agreement in number between verb and noun. When the VP consists of copula and adjective, an agreement in gender is added. All nouns in the two previous sentences and in the current clause before zp, which meet the morphological requirements, are candidates for antecedent.

After the test, there are three possibilities: 1) there is no noun; 2) there is exactly one noun; 3) there is more than one noun.

If there is no noun, the conclusion is that the antecedent is exophoric. But there are some cases when this is not true and we formulated two exceptions from this rule. The first exception concerns some nouns, which can be used with verb in singular or plural. The second concerns syntactic construction with verb in plural and a compound subject consisting of more than one noun in singular separated by comma or connected with conjunction. Formally, there is a disagreement because there is no noun in plural, but semantically there isn't.

When there is exactly one noun agreed with the verb, the conclusion is that this noun is the antecedent. Again, this is not always true. We have an exception to check the possibility for exophoric antecedent and for the syntactic construction, mentioned before.

Even when there is more than one noun, agreed with the verb, exophoric antecedent or the mentioned syntactic construction are possible. When these possibilities are rejected, the algorithm starts a procedure for finding the antecedent. A strategy, known as a list of preferences [12] can be applied. The preferences for Bulgarian are formulated and ordered after repeated analysis of the corpora. The main requirement to every preference is its universality, i.e. to be valid for as many as possible examples. During our work we formulated more heuristics, very useful in some cases, but harmful in others. Except the preferences themselves, their order is also very important, because the process stops when a noun satisfies one of them. Our list of preferences is:

1. The nearest personal pronoun in nominative or proper name, which precedes a verb and is not part of PP.
2. A noun already recognized as antecedent of the previous zp.
3. A noun before subordinate attribute clause.
4. A noun, which is not part of subordinate attribute clause.
5. A noun, which is not part of PP.
6. A noun before verb, which is not part of PP.

If after using these rules, there is still more than one candidate left, we use heuristics to choose the antecedent.

The analysis of this strategy, applied to the texts from corpora, shows some typical, recurring mistakes. In most of the cases the source of the mistakes are not the rules, but the strategy to take the first noun, satisfying the current rule and to reject the others. This approach is not successful in long sentences, where the attentional state moves gradually. Rules 1 and 2 are very strong and they will choose a wrong noun as an antecedent. The main disadvantages of this approach are that it does not render the change of attentional state and it takes into account just one feature of the candidate for antecedent. That is why our idea is to use more than one feature of the noun and to render the change of attentional state.

Five scores are assigned to all nouns which passed the test for morphological agreement. After that, these scores are changed according to the distance of the noun to the zp. The scores of the nouns from the previous clause are not changed, but the scores from the clause before previous are decreased by 0.2. The scores are decreased by 0.2 for every previous clause. As the distance to the zp increases, the scores decrease. For example, if the sentence consists of 4 clauses: W, X, Y, Z and in the final clause Z there is a zp, the scores of the nouns in clause Y will be 5. The scores of the nouns in clause X will be 4,8 and the scores of the nouns in clause W will be 4,6. After that the assigned scores are changed according to the syntax position of the noun.

Positive score of 0,4 gets a noun, which is an antecedent of already resolved zp.
Positive score of 0,3 get proper names and personal pronouns in nominative.
Positive score of 0,2 get nouns which have an attributive adjective.
Positive score of 0,2 get nouns before a verb.
Negative score of 0,1 get nouns which are part of PP.
Negative score of 0,3 get nouns which are part of subordinate attribute clause.

A noun, which fulfills several conditions, gets scores from all of them. This is not referring to already resolved antecedent. It always gets only 0,4 positive score. The noun with highest composite score is proposed as antecedent. If two or more nouns have equal scores, we have some heuristic rules to choose the antecedent.

VI. EVALUATION

Our research is oriented towards the algorithms for zero pronoun resolution, but not to the NLP system as a whole. The parser is not an object of the investigation. The texts to the input of the algorithm are correctly parsed, presented in XML format. Evaluation of the results is based on this presumption.

Two steps of the algorithm – marking the clauses with zero pronouns and choosing the antecedent are evaluated individually. We use precision (p), recall (r) and F-measure (F) as measures for the first part of the algorithm. By definition

$$p = \frac{Tp}{Tp + Fp} \qquad r = \frac{Tp}{Tp + Fn} \qquad F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Tp – true positives, Fp – false positives, Fn – false negatives.

The results, based on the analysis of our corpora are:

p = 87,7% r = 98,9% F = 93,46%

In [10] Mitkov argues that precision and recall are not suitable for evaluation of anaphora resolution algorithms. For evaluation of the two algorithms for zero pronoun resolution in Bulgarian we use the measures proposed by him: success rate, non-trivial success rate and critical success rate. By definition:

$$Success\ rate = \frac{Number\ of\ successfully\ resolved\ anaphors}{Number\ of\ all\ anaphors} = \frac{s}{n}$$

$$Non - trivial\ succes\ rate = \frac{s - k}{n - k}$$

k – number of those anaphors, which have only one candidate for antecedent.

$$Critical\ success\ rate = \frac{s - k - m}{n - k - m}$$

m – number of those anaphors, which are resolved on the basis of person and number agreement.

In the evaluation are also not included clauses with verbs in 1st or 2nd singular or plural, because the antecedent is obviously exophoric – the speaker or the hearer.

The results, based on the analysis of our corpora are:

First algorithm, based on list of preferences strategy:
Success rate = 85% Non-trivial success rate = 82% Critical success rate = 81%

Second algorithm, based on assigning scores strategy:
Success rate = 88% Non-trivial success rate = 85% Critical success rate = 84%

These results are satisfactory comparable to anaphora resolution algorithms for other languages. There is no other algorithm for Bulgarian and more concrete comparison cannot be done.

VII. FUTURE WORK

In our future work we will use the results of the syntactic parser created at the Department of Linguistic Modeling at Bulgarian Academy of Science. It will be available from the web. It will facilitate our work, relieving us from the process of creating the productions for the context free grammar. Our future work will be directed to machine learning approach for zero pronoun resolution in Bulgarian.

ACKNOLEDGEMENTS

The results of this study have been obtained within the framework of a project, funded by the Technical University of Sofia subsidy for scientific research in 2013.

REFERENCES

[1] Bulgarian Academy of Science, Institute of Bulgarian language, Grammar of the contemporary literary Bulgarian language, part III Syntax, BAS Publishing, Sofia 1983 (in Bulgarian)

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

[3] Grigorova D., Zero pronoun resolution in Bulgarian, Proceedings of CompSysTech'11, Vienna, Austria, pp. 399-404, ACM Vol. 578.

[4] Grigorova D., A rule based approach for zero pronoun resolution in Bulgarian, Proceedings of Sixth International Conference "Computer Science'11", Ohrid, Macedonia, CD, p.p. 106-111

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

[6] Halliday M. A. K., Hasan, R.: Cohesion in English, Longman, London, 1976

[7] Iida, R., Inui, K. and Matsumoto, Y., "Capturing salience with a trainable cache model for zero-anaphora resolution", Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/AFNLP-09), p. 647-655.

[8] 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.

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

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

[11] Pereira S., ZAC.PB: An Annotated Corpus for Zero Anaphora Resolution in Portuguese, Student Research Workshop, RANLP 2009 – Borovetz, Bulgaria, pp. 53-59

[12] 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

[13] Penchev Y., T. Boyadzhiev, I. Kucarov. Contemporary Bulgarian language. "Petar Beron" publishing company, Sofia 1998 (in Bulgarian)

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

[15] Rello L., I. Ilisei, A Comparative Study of Spanish Zero Pronoun Distribution
Available from: http://pers-www.wlv.ac.uk/~in0963/papers/ilisei_ISMTCL_2009.pdf

[16] Popov K., Contemporary Bulgarian language, Syntax, "Science and Art" publishing company, Sofia 1979 (in Bulgarian)

[17] Simov K., P. Osenova, A. Simov, M. Kouylekov. Design and implementation of the Bulgarian HPSG-based Treebank. In Erhard Hinrichs and Kiril Simov, editors, Journal of Research on Language and Computation, Special Issue, Kluwer Academic Publishers, pages 495-522

[18] Stoyanov St., Grammar of the contemporary literary Bulgarian language, Phonetics and Morphology, "Science and Art" publishing company, Sofia 1980 (in Bulgarian)

[19] Tanev Hr., R. Mitkov, Shallow language processing architecture for Bulgarian, Proceedings of the 19th International conference on Computational linguistics - Volume 1 Taipei, Taiwan, pp. 1-7, 2002. Available from: http://www.aclweb.org/anthology-new/C/C02/C02-1027.pdf

[20] Zhao, S. and Ng, H., "Identifcation and resolution of Chinese zero pronouns: a machine learning approach" Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CNLL-07), p.541-550.

ABOUT THE AUTHOR

Assist.Prof. Diana Grigorova, Department of Computer Systems, Technical University of Sofia, Phone: +359 89 5590213, E-mail: dgrigorova@tu-sofia.bg.