МЕТОДЫ РАЗРЕШЕНИЯ МЕСТОИМЕННОЙ АНАФОРЫ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Каменская М. А. (ma_kamenskaya@mail.ru)
Российский университет дружбы народов, Москва, Россия

Храмоин И. В. (hramoin@isa.ru),
Смирнов И. В. (ivs@isa.ru)
Институт системного анализа Российской академии наук, Москва, Россия

Ключевые слова: разрешение анафоры, машинное обучение, метод опорных векторов, деревья решений, семантические роли

DATA-DRIVEN METHODS FOR ANAPHORA RESOLUTION OF RUSSIAN TEXTS

Kamenskaya M. A. (ma_kamenskaya@mail.ru)
Peoples' Friendship University of Russia, Moscow, Russia

Khramoin I. V. (hramoin@isa.ru),
Smirnov I. V. (ivs@isa.ru)
Institute for Systems Analysis of RAS, Moscow, Russia

The paper considers two data-driven methods for anaphora resolution of Russian texts. These methods are based on machine learning with annotated corpora and using no additional information except linguistic features. The first method uses Support Vector Machine as learning and classifying algorithms, the second method uses Decision Tree inducer. We evaluate the performance of the methods with several feature sets and corpora. Feature sets included morphological, syntactic and semantic features. In this paper we also evaluate how semantic features, namely semantic roles, impact the performance of anaphora resolution in Russian. We used our manually annotated corpus as well as a corpus provided by the organizing committee of the forum for the evaluation of linguistic text analysis systems, an event of Dialogue 2014. Experiments showed that precision of SVM is higher on experimental data for almost all cases. It was shown that semantic features enhance the performance of the methods for anaphora resolution of Russian texts. We have also calculated the optimal distance between the anaphor and the hypothetic antecedent and used it in our methods.

Key words: anaphora resolution, machine learning, support vector machine, decision trees, semantic roles

1. Introduction

Anaphora resolution is one of the core problems of natural language processing. Methods for anaphora and coreference resolution are used in systems for machine translation, information retrieval, information extraction and others. The problem of anaphora resolution is widely researched for English and other European languages. For Russian this problem had been solving by different researchers but until Dialog-2014, there had been no objective evaluation of methods for anaphora resolution of Russian.

In this paper we solve two tasks: the first one is to evaluate two simple data-driven methods for anaphora resolution of Russian which use no additional information except linguistic features. These methods based on machine learning with several feature sets using annotated corpora. The second task is to investigate how semantic features, namely semantic roles, influence performance of anaphora resolution of Russian.

We deal only with pronominal anaphora resolution and compare two approaches—statistical one, based on Support Vector Machine, and inductive method for Decision Tree construction. As learning and testing data sets, we used two annotated corpora: the first is our own, the second one was provided by organizers of Dialog-2014 parsers evaluation task. Feature set included morphological, syntactic and semantic features, obtained from semantic parser developed in ISA RAS [Osipov et al., 2008].

In section 2 related works are reviewed, section 3 describes feature sets, section 4 describes methods and section 5 presents experiment results. Section 6 presents conclusion and future work.

2. Related works

Research in automatic pronominal anaphora resolution for English started in the 70th. The first methods and systems by Winograd, Wilks, Hobbs [Mitkov, 1999] operated with the rules relying mostly on syntactical information; in addition, encyclopedic knowledge was also widely applied. In the 80th the tendency of combining different features, which had been used separately before, appeared. The papers of E. Rich and S. LuperFoy, J. Carbonell, R. Mitkov described the algorithms that combined agreement of gender and number, syntactic and semantic relations. In the 90th rule-based algorithms were replaced with statistical data-driven algorithms. I. Dagan and A. Itai, Connolly, Burger used the machine learning methods for anaphora resolution for the first time. For learning more about works of that time, one can turn to the paper of R. Mitkov [Mitkov, 1999]. The author discusses the history of the problem, traditional methods for anaphora resolution and characterizes the known computer systems.

Modern approaches are based on automatic learning using annotated corpora. They combine traditional linguistic methods with statistical methods and use different types of knowledge: morphologic, syntactic, semantic, and additional information, such as thesauri. A lot of interesting ideas and methods were represented at the CoNLL-2011 Shared Task [Pradhan et al., 2011]. The system [Lee et al., 2011] based on combination of multi-pass sieves, which incorporate lexical, syntactic, semantic, and discourse information, showed the best results. A variety of data sets available

for learning coreference resolution systems for English (see, for example [MUC-6 data set, 1995]) provides progress of research in this field.

Anaphora resolution for Russian is less experimentally researched. In [Kibrik, 1996] author discusses theoretical aspects of the anaphora phenomenon for Russian language and describes the series of linguistic features, reflecting nature of anaphora. One of the latest Kibrik's papers [Kibrik et al., 2013] is rather informative. The works of Tolpegin [Tolpegin, 2006], [Tolpegin et al., 2006] are also well known. The author proposes algorithm for construction of statistic model for pronominal anaphora resolution in Russian texts using machine learning methods. In paper [Abramova et al., 2011] authors describe in detail principles of anaphoric relations detection in different sentences and situations, which they use for the analysis of rules of socio-political texts coherence. The research of Mal'kovskij [Mal'kovskij et al., 2013] is one from the latest known papers for Russian. The work deals with the rule-based method for pronominal anaphora resolution, which uses analysis of words collocation. The core problem for Russian is absence of open data sets for learning coreference resolution methods and their evaluation.

There are the series of works that evaluate the influence of semantic knowledge on anaphora resolution quality. The papers [Ponzetto and Strube, 2006], [Kong et al., 2008], [Huang et al., 2009], [Zhou et al., 2001] demonstrate that using semantic roles as additional features improves anaphora resolution performance for English. The authors use data-driven methods but the approaches to selecting the groups of features and learning algorithms are different.

3. Feature set

We consider anaphora resolution problem as classification problem and solve it using machine-learning methods. The following features were used for leaning and classification:

Morphological and syntactic features:

1) gender, number, case, and animate of anaphor;
2) gender, number, case, and animate of antecedent;
3) comparison of anaphora's animate and antecedent's animate;
4) number of sentences between anaphor and antecedent;
5) number of words between anaphor and antecedent;
6) number of hypothetical antecedents between anaphor and antecedent;
7) number of nouns between anaphor and antecedent;
8) name of syntactic relation between anaphor and antecedent;

Semantic features:

9) semantic roles of anaphor;
10) semantic roles of antecedent;
11) combination of categorical semantic class of the head word of syntactic phrase, which contains anaphor as related word, and categorical semantic class of the head word of syntactic phrase, which contains antecedent as related word;

12) combination of categorical semantic class of the head word of syntactic phrase, which contains anaphor as related word, and categorical semantic class of the antecedent.

We use features 1–3, because we suppose that gender, number and animate of anaphor should agree with gender, number and animate of antecedent. Features 4–7 give information about distance between anaphor and antecedent in different scales. Feature 8 was proposed, because we guess that antecedent can be a part of fixed number of syntactic relations as a related word. We also expect antecedent to be a related component in verbal phrase. A word can be labeled with several semantic roles, because it can be an argument for different situations described in one sentence, especially in complex sentences. We use features 11–12, because we suppose that the categorical semantic class of noun can be combined with the fixed number of categorical semantic classes of verbs, as well as categorical semantic classes of verbs, which are associated with the same noun, are combined according to the special rules.

Features' values were obtained as a result of morphological, syntactic and semantic analysis of texts [Osipov et al., 2008]. Methods for semantic role labeling of Russian are described in [Smirnov et al., 2014]. Detailed lists of categorical semantic classes and semantic roles are presented in [Osipov, 2001]. We experimented with two feature sets: feature set FS-1 included features 1–8, feature set FS-2 included features 1–8 and semantic features 9–12.

4. Methods

Anaphora resolution is a task of detecting correct pairs "antecedent-anaphor". In our research, we deal only with personal, reflexive and demonstrative pronouns. The training set contains examples of correct and incorrect "antecedent-anaphor" pairs. Correct "antecedent-anaphor" pair contains correct hypothetic antecedent, incorrect "antecedent-anaphor" pair contains incorrect hypothetic antecedent. Hypothetic antecedent is a noun or pronoun for which anaphora has been already resolved. Hypothetic antecedent must be agreed with anaphor by number and gender. The distance in words between the anaphor and the antecedent should be not more than preliminary defined value that depends on corpus. Training example is presented as set of values of named features described in the previous section.

4.1. The algorithm of constructing training data set using annotated corpus

1. Find first annotated "antecedent-anaphor" pair in corpus.
2. Look for all nouns or pronouns for which anaphora has been already resolved, between the anaphor and the antecedent. Their number and gender must be agreed with anaphor's number and gender. The search area is limited by a predefined number of words.

3. All nouns and pronouns, which were found in the step 2 are incorrect hypothetic antecedents.
4. If correct antecedent is not in a search area, it will not be added to the training set.
5. Do steps 1–4 for each annotated example.

For training and classifying correct/incorrect pairs we used Support vector machine method (SVM) [Chang and Lin, 2014] and decision tree method [University of Waikato, 2014] with REPTree learner.

4.2. The algorithm for anaphora resolution

1. Find first anaphor, for which antecedent has not already been found. If anaphor has not been found, algorithm finishes.
2. Look for all nouns or pronouns for which anaphora has been already resolved, between the anaphor and the antecedent. Their number and gender must be agreed with anaphor's number and gender. The search area is limited by a predefined number of words.
3. Add them to hypothetic antecedents' set.
4. Assign to each pronoun in hypothetic antecedents' set categorical semantic class of its' antecedent.
5. Calculate the probability of each hypothetic antecedent to be correct antecedent using classification method.
6. Choose antecedent which has the highest probability and link it with the concerned anaphor. Go to step 1.

Area for searching hypothetic antecedent is limited by the number of words in step 2, because anaphor usually refers to nearest hypothetic antecedent. This value has been calculated in our experiments.

5. Results of experiments

Experiments have been run on several manually annotated corpora. The first corpus CORPUS-1 contains 17 texts of the Moshkov's library and 34 texts of SynTagRus [Apresjan et al., 2005]. CORPUS-1 contains 910 "antecedent-anaphor" pairs. CORPUS-2 is the annotated corpus, provided as a training set by the organizing committee of the Dialogue-2014 forum for the evaluation of linguistic text analysis systems. CORPUS-2 contains 92 texts and 967 "antecedent-anaphor" pairs. CORPUS-3 is union of CORPUS-1 and CORPUS-2.

The results of the preliminary experiments showed that distance in words, which limit hypothetic antecedent's searching area, is one of the most important features. This feature has a significantly positive effect on precision and recall of automated anaphora resolution because rise in distance between hypothetic antecedent

and anaphor lowers probability, that the anaphor refers to this antecedent [Tolpegin et al., 2006]. Moreover, rise in that distance causes rise in number of hypothetic antecedents, which makes anaphora resolution more complicated, costly and long. This is the reason to find the optimal distance that limits hypothetic antecedent's searching area. Such distance should include correct antecedent most time and limit the number of hypothetic antecedents as much as possible.

The distributions of number of correct "antecedent-anaphor" pairs according to distance between antecedent and anaphor for each corpus are presented on figures 1–3. We calculated the optimal distance that covers 90% of correct "antecedent-anaphor" pairs for every corpus, using these distributions. This optimal distance is equal to 25 words for CORPUS-1, 14 words for CORPUS-2 and 18 words for CORPUS-3. Hypothetic antecedents were searched not farther than calculated optimal distance in both training and classifying process.

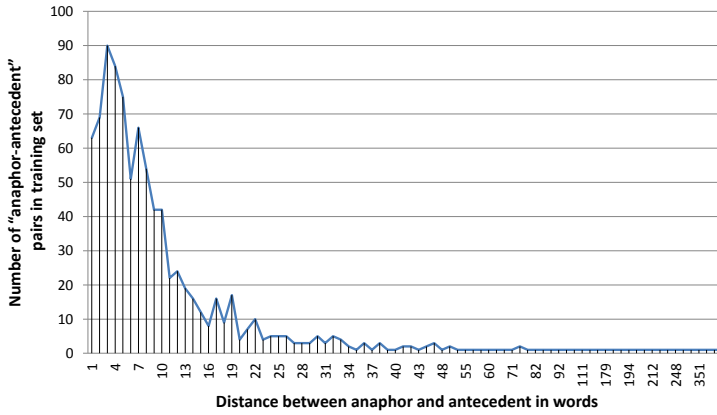

Fig. 1. Number of "antecedent-anaphor" pairs in relation on distance between anaphor and antecedent in CORPUS-1

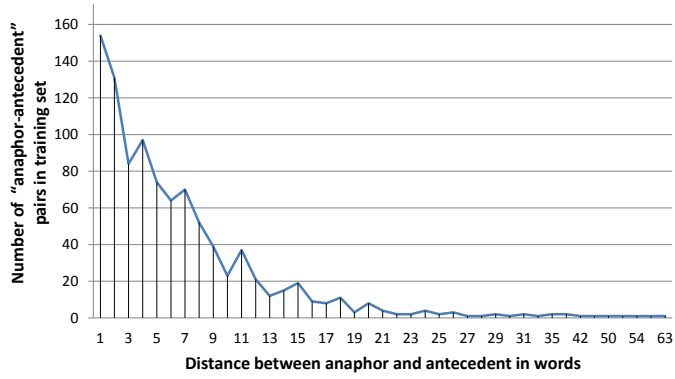

Fig. 2. Number of "antecedent-anaphor" pairs in relation on distance between anaphor and antecedent in CORPUS-2

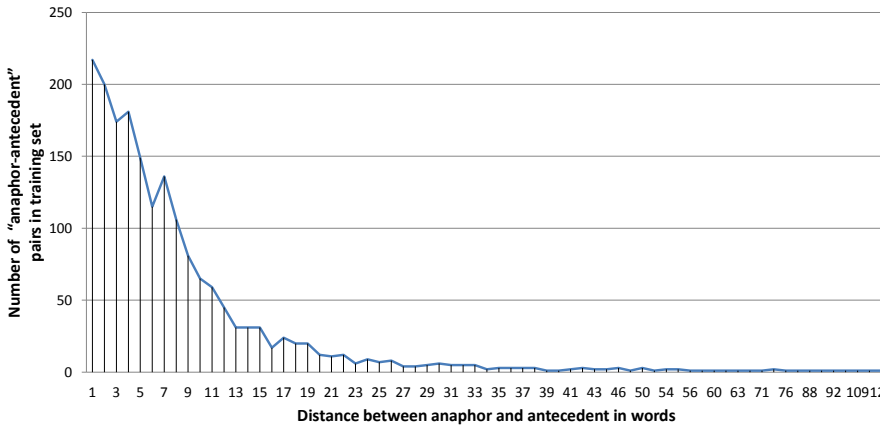

Fig. 3. Number of "antecedent-anaphor" pairs in relation on distance between anaphor and antecedent in CORPUS-3

We used the following metrics: SCORE-1 is a precision of recognition of both correct and incorrect "antecedent-anaphor" pairs (precision of classification of examples into two classes—correct or incorrect), SCORE-2 is the precision of finding correct antecedent for each anaphor. We use only CORPUS-2 as testing corpus to calculate SCORE-2. SCORE-2 represents actual precision of anaphora resolution and is most close to the task. We use ten-fold cross validation to calculate SCORE-1. Scores of SVM method with feature set FS-1 were chosen as a baseline.

The result of the first experiment on CORUS-1 is presented in table 1.

Table 1. Precision of anaphora resolution for different methods and feature sets on CORPUS-1 as training corpus

Feature set	SVM	REPTree
SCORE-1		
FS-1	0.811	0.773
FS-2	0.821	0.789
SCORE-2		
FS-1	0.473	0.484
FS-2	0.539	0.529

The result of the second experiment on CORUS-2 is presented in table 2.

Table 2. Precision of anaphora resolution for different methods and feature sets on CORPUS-2 as training corpus

Feature set	SVM	REPTree
SCORE-1		
FS-1	0.746	0.746
FS-2	0.771	0.747
SCORE-2		
FS-1	0.603	0.592
FS-2	0.61	0.609

The result of the third experiment on CORUS-3 is presented in table 3.

Table 3. Precision of anaphora resolution for different methods and feature sets on CORPUS-3 as training corpus

Feature set	SVM	REPTree
SCORE-1		
FS-1	0.766	0.634
FS-2	0.781	0.689
SCORE-2		
FS-1	0.571	0.548
FS-2	0.579	0.553

We have done 8 runs on the test corpus provided by the organizing committee of the Dialogue-2014 forum for the evaluation of linguistic text analysis systems. The methods were learned on CORPUS-3 with feature sets FS-1 and FS-2.

6. Conclusion and future work

The results of experiments showed the reasonable results for both simple methods. SVM exceeded DT by 0,1%–13,2% of precision for almost all runs on experimental data. Anaphora resolution using semantic features showed precision gain by 0,1%–6,6% for all methods and runs. Such values for precision gain in this case can be explained by the fact that semantic roles were assigned to both anaphor and antecedent in only 8% of anaphoric pairs in manually annotated learning corpus. The F-measure of semantic parser used for testing anaphora resolution is 75% with 67% of recall and 86% of precision, so precision gain for anaphora resolution is adequate and rather good. The best result on test corpus (61% of precision) was shown by SVM on CORPUS-2 with feature set FS-2.

Thus, experiments showed that semantic features enhance performance of methods for pronominal anaphora resolution of Russian texts. As a future work, we will extend feature set with extra-lingual information using several thesauri and enhance method for identifying hypothetic antecedents.

References

1. *Abramova N. N., Abramov V. E., Nekrasova E. V., Ross G. N.* (2011), Statistic analysis of social and political texts coherence [Statisticheskij analiz svjaznosti tekstov po obshchestvenno-politicheskoj tematike], Proceedings of the 13th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" [Trudy 13j Vserossijskoj nauchnoj konferentsii "Èlectronnye biblioteki: perspektivnye metody i tehnologii, èlektronnye kollektsii"], Voronezh, pp. 127–133.
2. *Apresjan J. D., Boguslavskij I. M., Iomdin B. L., Iomdin L. L., Sannikov A. V., Sannikov V. G., Sizov L. L.* (2005), Syntactically and semantically annotated corpus of Russian language: Present state and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo jazyka: sovremennoe sostojanie i perspektivy], National Corpus of Russian Language: 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], pp. 193–214.
3. *Chang C.-C., Lin C.-J.* (2014), LIBSVM—A Library for Support Vector Machines, available at: www.csie.ntu.edu.tw/~cjlin/libsvm/
4. *Huang Z., Zeng G., Xu W., Celikyilmaz A.* (2009), Accurate semantic class classifier for coreference resolution, Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 1232–1240.
5. *Kibrik A. A.* (1996), Anaphora in Russian narrative discourse: A cognitive calculative account In B, Fox (ed.) Studies in anaphora, Amsterdam, pp. 255–304.
6. *Kibrik A. A., Dobrov G. B., Khudyakova M. V., Loukachevitch N. V., Pechenyj A.* (2013), A corpus-based study of referential choice: Multiplicity of factors and machine learning techniques, Text processing and cognitive technologies. Cognitive modeling in linguistics: Proceedings of the 13th International Conference, Corfu, pp. 118–126.
7. *Kong F., Li Y., Zhou G., Zhu Q., Qian P.* (2008), Using Semantic Roles for Coreference Resolution, International Conference on Advanced Language Processing and Web Information Technology, pp. 150–155.
8. *Lee H., Peirsman Y., Chang A., Chambers N., Surdeanu M., Jurafsky D.* (2011), Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* (CONLL Shared Task '11), Stroudsburg, pp. 28–34.
9. *Mal'kovskij M. G., Starostin A. S., Shilov I. A.* (2013), Method of pronominal anaphora resolution in parallel with syntactic analysis [Metod razreshenija mestoimennoi anafory v protsesse sintaksicheskogo analiza], Perspective innovations in science, education, production and transport'2013 [Perspektivnye innovatsii v nauke, obrazovanii, proizvodstve I transporte'2013], available at: www.sworld.com.ua/index.php/ru/technical-sciences-413/informatics-computer-science-and-automation-413/20828-413-0615.

10. *Mitkov R.* (1999) Anaphora resolution: the state of the art, Working paper (based on the COLING'98/ACL'98 tutorial on anaphora resolution), available at: clg.wlv.ac.uk/papers/mitkov-99a.pdf
11. MUC-6 data set, (1995), available at: http://cs.nyu.edu/faculty/grishman/muc6.html
12. *Osipov G. S., Smirnov I. V., Tikhomirov I.* (2008), Relational–situational method for search and analysis of texts and its applications [Reljatsionno-situatsionnyj metod poiska i analiza tekstov i ego prilozhenija], Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], (1), pp. 3–10.
13. *Osipov G. S.* (2011), Methods of artificial intelligence [Metody iskusstvennogo intellekta], FIZMATLIT, Moscow.
14. *Ponzetto S. P., Strube M.* (2006), Semantic role labeling for coreference resolution, Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (EACL'06), Trento, pp. 143–146.
15. *Pradhan S., Ramshaw L., Marcus M., Palmer M., Weischedel R., Xue N.* (2011), CoNLL-2011 shared task: modeling unrestricted coreference in OntoNotes, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* (CONLL Shared Task'11), Stroudsburg, pp. 1–27.
16. *Smirnov I. V., Shelmanov A. O., Kuznetsova E. S., Hramoin I. V.* (2014), Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts [Semantiko-sintaksicheskij analiz estestvennyh jazykov Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov], Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], (1), pp. 95–108.
17. *Tolpegin P. V.* (2006), The new methods and algorithms of automated third person pronominal reference resolution of Russian texts, [Novye metody i algoritmy avtomaticheskogo razreshenija referentsii mestoimenij tret'ego litsa russkojazychnyh tekstov], Komkniga, Moscow.
18. *Tolpegin P. V., Vetrov D. P., Kropotov D. A.* (2006), Automated third person anaphora resolution algorithm on the basis of machine learning methods [Algoritm avtomatizirovannogo razreshenija anafory mestoimenij tret'ego litsa na osnove metodov mashinnogo obuchenija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006", [Komp'juternaja Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2006"], Bekasovo, pp. 504–507.
19. University of Waikato, (2014), Weka 3: Data Mining Software in Java, available at: www.cs.waikato.ac.nz/ml/weka/
20. *Zhou H., Li Y., Huang D., Zhang Y., Wu C., Yang Y.* (2011), Combining syntactic and semantic features by SVM for unrestricted coreference resolution, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* (CONLL Shared Task'11), Stroudsburg, pp. 66–70.