Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2017”

Moscow, May 31—June 3, 2017

COMPLEX APPROACH TOWARDS ALGORITM LEARNING FOR ANAPHORA RESOLUTION IN RUSSIAN LANGUAGE

Gureenkova O. A. (ol.gure@gmail.com)[1],
Batura T. V. (tatiana.v.batura@gmail.com)[2,3],
Kozlova A. A. (noriel266@gmail.com)[2],
Svischev A. N. (alekseisvischev@gmail.com)[2]

[1]Expasoft Ltd., Novosibirsk, Russia; [2]Novosibirsk State University, Novosibirsk, Russia; [3]A. P. Ershov Institute of Informatics systems, Novosibirsk, Russia

The paper considers applying of ensemble algorithm based on rules and machine learning for anaphora resolution in Russian language. Ensemble presents combination of formal rules, a machine learning algorithm Extra Trees and an algorithm for working with imbalanced learning sets Balance Cascade. Complexity of the approach lies in generation of complex features from rules and vectorization of syntactic context, with context data obtained from algorithms mystem (Yandex), SyntaxNet (Google) and Word2Vec.

Key words: anaphora, antecedent, cataphora, Random Forest, machine learning, imbalanced set, Extra Trees, Balance Cascade, SyntaxNet, Word2Vec

КОМПЛЕКСНЫЙ ПОДХОД К ОБУЧЕНИЮ АЛГОРИТМОВ ДЛЯ РАЗРЕШЕНИЯ АНАФОРЫ В РУССКОМ ЯЗЫКЕ

Гureенкова О. А. (ol.gure@gmail.com)[1],
Батура Т. В. (tatiana.v.batura@gmail.com)[2,3],
Козлова А. А. (noriel266@gmail.com)[2],
Свищев А. Н. (alekseisvischev@gmail.com)[2]

[1]ООО «Экспасофт», Новосибирск, Россия;
[2]Новосибирский государственный университет, Новосибирск, Россия; [3]Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия

В работе рассматривается применение ансамбля алгоритмов, основанного на правилах и машинном обучении для разрешения анафоры в русском языке. Ансамбль представляет собой объединение формальных правил, алгоритма машинного обучения Extra Trees и алгоритма для работы с несбалансированной выборкой Balance Cascade. Комплексность подхода заключается в генерации сложных признаков, полученных на основе правил и векторизации синтаксического контекста с учетом данных из алгоритмов mystem (Yandex), SyntaxNet (Google) и Word2Vec.

Ключевые слова: анафора, антецедент, катафора, случайный лес, машинное обучение, несбалансированная выборка, Extra Trees, Balance Cascade, SyntaxNet, Word2Vec

Introduction

In such natural language processing tasks as machine translation, information extraction and others the engineers often face the problem of anaphora resolution. Resolution of personal pronoun anaphora is the task of finding a word expression that a personal pronoun refers to. There is a significant number of researches related to anaphora resolution in the European languages [2, 3, 5]. As for the Russian language, the problem is not sufficiently represented. This is due to the fact that there is a lack of open annotated Russian corpora that are required for model training and evaluation.

The basic concepts related to the task of pronominal anaphora resolution are anaphor and antecedent. Consider the example: "***Человек*** *ленив по своей природе, и только жесточайшая конкуренция может привести* ***его*** *к успеху.*" The word "*его*", anaphor, refers to the same real-world entity that "*человек*", antecedent. Commonly, the antecedent is located in the text before the anaphor. But there are also cases of *cataphora*—a phenomenon, opposite to anaphora, when the antecedent appears in the text after the pronoun. In this work we aggregate these two terms into the one—anaphora.

The aim of our work was to develop an algorithm that resolves anaphoric links between the personal pronouns and their antecedents in the Russian language. In this algorithm we used a hybrid approach, based on rules and machine learning.

The anaphora resolution in Russian involves certain difficulties. Usually text preprocessing includes the following steps: part-of-speech tagging, morphological analysis of words, detection of noun phrases, syntactic parsing and surface-semantic analysis. During automatic preprocessing the errors tend to appear and accumulate at the following steps, and later the errors may affect the algorithm quality. However, interest in the problem of anaphora resolution remains high in recent years among both European and Russian researchers.

1. Related works

We can distinguish three approaches to anaphora resolution: rule-based, based on machine learning and hybrid.

A rule-based approach is suggested in [3]. The main idea is to take into account, besides part-of-speech tagging, the information about the noun phrases preceding the anaphora at the distance of two sentences. Only those noun phrases that agree with the anaphora in gender and number are selected. Then the rules are applied consistently. The size of text corpus for testing the algorithm was 28 thousand words (among which 422 pronouns). The accuracy was 57%.

The paper [1] describes a rule-based method of anaphoric links detection for the Russian language. The research is mainly aimed at studying the types of substitution used in various socio-political texts. Unfortunately, the authors did not provide a comparison of the accuracy of anaphoric relations detection.

Some researches [6, 9, 10] propose to solve the problem of anaphora detection using machine learning methods. In particular, the authors of the work [9] observed that if the support vector machine (SVM) is used in addition to a set of rules, then the best accuracy is 52.04%. The study [6] found that additional knowledge about the semantic roles of anaphora and antecedent can improve the quality of the solution of the problem by 0.1–6.6%.

The study [2] describes pronominal anaphora resolution in analysis of user opinion data. The authors used 16 characteristics, divided into three categories: anaphoric pronouns, candidates for antecedents and relationship characteristics. A relatively small corpus in the Basque language was taken for training and testing. It consisted of 50 thousand words and 249 anaphoric pronouns. Various methods of machine learning were compared in the experiment: support vector method (SVM), ensemble of decision trees (RF), k-nearest neighbors (kNN) method, multilayer perceptron (MLP), Bayes method (NB), Bayesian combined approach and Decision trees (NB-Tree). Quality assessment was carried out using 10-fold cross-valuation. The results of the experiments showed that a high accuracy of 0.803 is observed for the SVM, while the best recall of 0.702 and the F-measure of 68.3% were obtained using the RF.

An article [7] describes the experiment on the anaphora resolution for the Russian language using a hybrid approach. First, a set of potential antecedents is selected for each pronoun. Next, the most likely candidate is chosen on the basis of a set of characteristics containing information on compatibility of words, statistical, morphological and syntactic characteristics. After that the Random Forest algorithm is used for classification of feature vectors. The highest accuracy of 71% was obtained on the set of all available features.

A hybrid approach to coreference resolution in the English language is presented in [5]. The authors proposed 10 models based on the rules as features for machine learning. For example, the rule “Is there an anaphor-antecedent pair in direct speech?” can be considered as a binary categorical feature. Some of the rules appeared in the article [5] were applied in our work.

The method proposed in our paper is based on machine learning and implementation of a complex approach to feature matrix generation. The feature matrix contains features obtained from the rules or generated from other features. To analyze the syntactic context, a neural network algorithm SyntaxNet [8] was used.

2. Data preparation

We used a text corpus1 of 2,684 texts on criminalistic topics from the informational portal mvd.ru to train our model. The mean text length is about 1,200 symbols. 1,000 more texts were taken for testing. The texts were annotated manually by expert linguists.

Required data were extracted from the texts and transformed into a feature matrix. The matrix rows are represented by all possible pronoun-noun pairs of a single text, some of which are correct anaphoric pairs. The correct pairs got the positive class labels according to the annotation. Thus, the anaphora resolution task was reduced to the binary classification of pronoun-noun pairs.

In the first experiments we searched the antecedents for current pronoun throughout the whole text. But it was founded out that it is rather meaningless and moreover, requires a very large amount of computational resources. The probability of finding an antecedent far away from the pronoun is too low to consider such cases. The experiments described in [9] showed that the optimal window for searching the antecedents is 23 words. Given that the average sentence length in Russian is roughly equal to 10 words, we decided to limit the window to two sentences before and two sentences after the sentence with antecedent. The antecedents that appear after the pronoun are the cases of cataphora. Cataphoric pairs accounted for 33% of all data. The size of training sample was 262,804 pairs pronoun-noun.

Another positive effect of such restriction was partial solution of the imbalanced set problem. The percentage of correct anaphoric pairs was 3% before the limitation, and it increased to 10% after setting a window.

3. Feature Generation and Selection

The feature matrix contains features of anaphor and antecedents which are based on morphological, syntactic, statistical and vector analysis of texts. The whole number of generated features was 2,596, but after the selection only 240 features left.

Feature selection was semi-automatic. The features were ranked by their importance, evaluated by the Extra Trees model, which used for classification. Moreover, the Recursive Feature Elimination (RFE) method was used. The features were divided into several groups, then random features were sequentially removed from each group, and the quality of classification with the current feature group was estimated.

It does not seem possible to describe separately all the used features in the scope of this article because of their large number, but it is possible to combine features into the following groups (see Table 1).

---

[1] This corpus is available at https://github.com/my-master/CoreferenceData

Table 1. Feature groups of training set

Group description	Feature origin	Number of features
1. *Binary categorical features*, obtained with mystem morphological analyzer and relatively complex rules (for example, "the entity indicated by the antecedent is a person", "anaphor and antecedent agree in person, number and case").	rules, mystem	7
2. *Non-binary categorical features*. They are based on simple grammatical characteristics of anaphor and antecedent (part of speech, number, gender, case, animacy, type of anaphoric pronoun, syntactic relations).	mystem, Syntaxnet	82 (after binarization)
3. *Numerical features* derived from Word2Vec vectors for syntactic contexts. They include all possible distances between context vectors of antecedent and anaphor and distances between antecedent and anaphor own vectors and average vectors of their contexts.	Word2Vec, SyntaxNet	28
4. *Numerical features*, obtained as a result of calculating various linear (i.e., not syntactic) distances, for example, distance in words, sentences, nouns, verbs between anaphor and antecedent.	rules, mystem	13
5. *Transposed vectors*, obtained using SyntaxNet with TF-IDF vectors of morphological and syntactic tags, taken for anaphor and antecedent in three directions of the syntactic context: the child nodes, the parent node, and the sibling nodes.	SyntaxNet	110

4. Classification Process

We considered the problem of anaphora resolution as a binary classification problem of possible pronoun-noun pairs. In view of the large space dimension and the variety of ways to obtain features, it was decided to use the algorithm based on decision trees as the classification algorithm. It was revealed that the Extra Trees [4] algorithm, which is a modification of a random forest, shows the best results. Therefore, it was chosen as the main algorithm.

For the Extra Trees algorithm, the following parameters were selected:

- the maximum percentage of features for finding the best partition was 0.23;
- the number of trees in the forest was 200;
- balanced class weighting method was chosen.

In addition to choosing the main algorithm, it was also necessary to solve the problem with an imbalanced sample, since after the initial screening of incorrect anaphoric pairs by a three-sentence frame, the proportion of correct pairs was still at 10%, which

could lead to low accuracy of the trained algorithm. To solve this problem, various simple methods were tested (reducing the number of objects of the major class, duplication of the objects of the minor class), which did not bring a gain in quality. Nevertheless, after applying the ensemble algorithm Balance Cascade F-measure improved by 1%.

The following parameters for the Balance Cascade were selected:

- fraction of the minor class: 0.5;
- maximum number of generated sub-samples: 200;
- random-forest was chosen as internal classifier for quality assessment.

After applying the Balance Cascade algorithm, new true anaphoric pairs were generated and some of the old incorrect pairs were discarded. As a result, the proportion of true pair from the entire sample has already become 25%. However, due to the specificity of the Balance Cascade algorithm, the size of the training sample increased approximately 1.3 times, which increased the requirements for computing power. For example, before the application of Balance Cascade, the size of the training matrix was $262{,}804 \times 240$. After converting the sample, the matrix size varies from $341{,}900 \times 240$ to $345{,}800 \times 240$.

5. Experiment results

Precision, recall and F-measure were used to assess the quality of the proposed method. It is necessary to take into account both precision and recall simultaneously for evaluating the results. Fig. 1 shows the Precision-Recall curve.

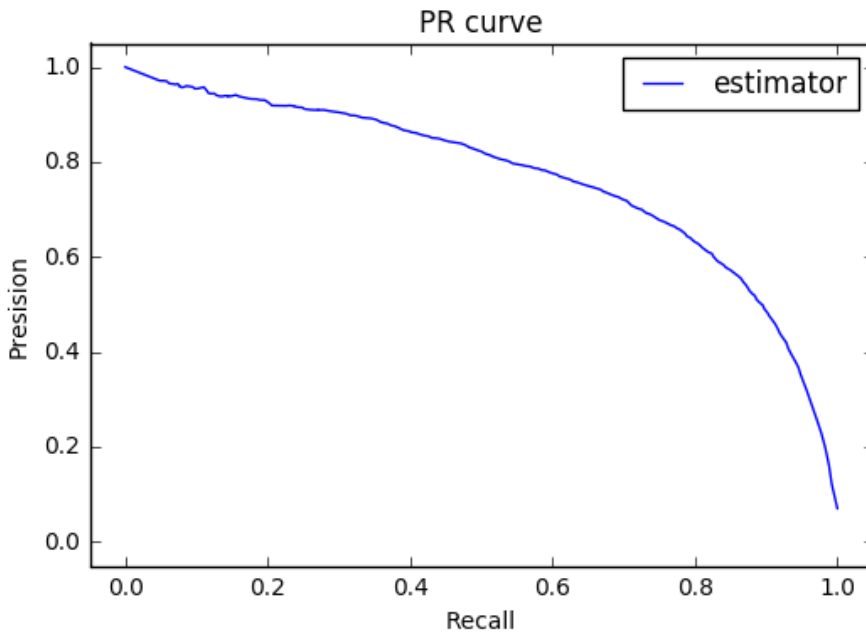

Fig. 1. Precision-Recall curve

Accuracy was not taken into account, since under the conditions of an extremely imbalanced sample it would be high even with a constant classifier that assigns the value of the wrong pair to all pairs. Table 2 gives the best obtained values of precision and recall for a certain threshold of the probability of belonging to the class of correct anaphoric pairs.

Table 2. Precision and recall values obtained in the control sample

Threshold of the probability	Precision	Recall
0.280	0.6577	0.7789
0.285	0.6605	0.7748

Due to high relevance of the feature groups it was decided to test the algorithm quality on each group and their combinations separately. Also it helped to understand the contribution of each group to the overall result. The F-score obtained on different feature groups is shown at the Table 3.

Table 3. F-score on feature groups

Feature group	F-score, %
Binary categorical features	28.6
Non-binary categorical features	57.8
Numerical features derived from Word2Vec vectors	50.0
Numerical features (linear distances)	32.9
Transposed vectors	61.2
Binary categorical features *Non-binary categorical features* *Numerical features derived from Word2Vec vectors* *Numerical features (linear distances)*	70.3
Binary categorical features *Non-binary categorical features* *Numerical features derived from Word2Vec vectors* *Transposed vectors*	70.2
Non-binary categorical features *Numerical features derived from Word2Vec vectors* *Numerical features (linear distances)* *Transposed vectors*	70.1
Binary categorical features *Non-binary categorical features* *Numerical features (linear distances)* *Transposed vectors*	70.9
Binary categorical features *Non-binary categorical features* *Numerical features derived from Word2Vec vectors* *Numerical features (linear distances)* *Transposed vectors*	**71.4**

It can be seen that in the cases when only one of the feature groups was taken into account, the corresponding F-scores differ greatly from each other. The best value of 61.2% is achieved on the transposed vectors obtained using SyntaxNet with TF-IDF vectors. Presumably this is due to the fact that the fifth feature group is the most numerous, i.e. we can see that there is a correlation between the number of features in each group and the obtained result.

At the same time, in the cases when different combinations of feature groups are used simultaneously, their corresponding F-measures differ insignificantly. It implies that despite the correlation among the features, decreasing their number doesn't lead to increase of the F-measure. The best value of 71.4% was obtained in the case when all five feature groups were used.

6. Conclusion

In this article, we offer a complex approach to the anaphora resolution in the Russian language. Formally, the problem of anaphora resolution can be represented as a binary classification problem. The feature matrix for classification contains information about morphological, syntactic, statistical and vector analysis of texts. The total number of generated features was 2,596, but after the selection only 240 the most important features left. All the features can be divided into five groups. Despite the correlation among the features, decreasing their number does not lead to increase of the F-measure.

Acknowledgment

The work was carried out at the Novosibirsk State University with the financial support of the Ministry of Education and Science of the Russian Federation (contract No. 02.G25.31.0146) as part of the implementation of RF Government Decision No. 218 "On State Support Measures for the Development of Cooperation between Russian Higher Educational Institutions and Organizations Implementing Comprehensive Projects for High-Tech Production".

References

1. *Abramov V., Abramova N., Nekrasova E., Ross G.* (2011), Statistical Analysis of the Coherence of Texts on Social and Political Issues [Statisticheskij analiz svjaznosti tekstov po obshhestvenno-politicheskoj tematike], Proceedings of the 13th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections RCDL'2011" [Trudy 13j Vserossijskoj nauchnoj konferencii «Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii» — RCDL'2011], Voronezh, Russia, pp. 127–133.
2. *Arregi O., Ceberio K., Díaz de Illarraza A., Goenaga I., Sierra B., Zelaia A.* (2010), Determination of Features for a Machine Learning Approach to Pronominal Anaphora Resolution in Basque, Procesamiento del language natural, N 45, pp. 291–294.

3. *Barbu C., Mitkov R.* (2001), Evaluation tool for rule-based anaphora resolution methods, Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 34–41.
4. *Geurts P., Ernst D., Wehenkel L.* (2006), Extremely randomized trees, Machine Learning, Vol. 63, N 1, pp. 3–42.
5. *Jurafsky D., Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M.* (2013), Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules, Association for Computational Linguistics, Vol. 39, N 4, pp. 885–916.
6. *Kamenskaya M. A., Khramoin I. V., Smirnov I. V.* (2014), Data-driven Methods for Anaphora Resolution of Russian, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014), Issue 13 (20), pp. 241–250.
7. *Malkovskiy M. G., Starostin A. S., Shilov I. A.* (2013), Method of pronoun anaphora resolution in parallel with syntactic analysis [Metod razreshenija mestoimennoj anafory v processe sintaksicheskogo analiza], Collection of scientific works SWorld on materials of the international scientific-practical conference [Sbornik nauchnyh trudov SWorld po materialam mezhdunarodnoj nauchno-prakticheskoj konferencii], Vol. 11, N 4, pp. 41–49.
8. *Petrov S.* (2016), Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source, available at: https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html
9. *Protopopova E. V., Bodrova A. A., Volskaya S. A., Krylova I. V., Chuchunkov A. S., Alexeeva S. V., Bocharov V. V., Granovsky D. V.* (2014), Anaphoric Annotation and Corpus-Based Anaphora Resolution: An Experiment, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014), Issue 13 (20), pp. 562–571. URL: http://www.dialog-21.ru/media/1125/dialogue2014_full_version.pdf
10. *Tolpegin P. V.* (2008), Automatic coreference resolution of third person pronouns in Russian texts [Avtomaticheskoe razreshenie koreferencii mestoimenij tret'ego lica russkojazychnyh tekstov], Theses for the degree of candidate of technical sciences, Moscow, 241 p.