THE IMPACT OF MORPHOLOGY PROCESSING QUALITY ON AUTOMATED ANAPHORA RESOLUTION FOR RUSSIAN

Ionov M. (m.ionov@corp.mail.ru)
Mail.ru Group, Moscow, Russia

Kutuzov A. (andrey.kutuzov@corp.mail.ru)
Mail.ru Group, National Research University Higher School of Economics, Moscow, Russia

The paper deals with the problems of creating and tuning a system of automated anaphora resolution for Russian. Such a system is introduced, combining rule-based and machine learning approaches. It shows F-measure from 0.51 to 0.59. Freeling serves as an underlying morphological layer and an account of its quality is given, with its influence on anaphora resolution workflow. The anaphora resolution system itself is available to download and use, coming with online demo.

Keywords: anaphora resolution, morphology processing, machine learning, antecedent

1. Introduction

In this paper we describe **An@phora**—the system for automated pronominal anaphora resolution in Russian texts. The system was built as a participant of anaphora resolution systems evaluation forum to be held at the conference 'Dialog—2014'. It combines rule-based and machine learning approach to achieve better quality.

Anaphora (and coreference in general) resolution is crucial to many natural language processing applications, including dialog agents, machine translation, question answering systems, and a lot more. At the same time, Russian natural language processing community lacks open tools to accomplish this task. Unfortunately, published reports on automated Russian anaphora resolution (see the section 2) are few and most of the time not extensively documented with regard to precision and recall of approaches used. What's even more disappointing, tools themselves are not published. To our knowledge, up to now there is no open and available system for Russian anaphora resolution.

Thus, we addressed not only the task of constructing anaphora resolution machine itself, but also of making it publicly available under an open-source license. The system is implemented in Python, and is free to download and use[1]. Also, live demo is available[2], using Brat on-line markup system [Stenetorp et al 2012].

---

[1] https://github.com/max-ionov/russian-anaphora

[2] http://ling.go.mail.ru/anaphora

At the same time, it should be stressed that as in many other areas of natural language processing, quality of underlying stages of linguistic analysis is crucial for performance of the system. In the case of anaphora resolution, of great importance are tools which provide tokenization, sentence splitting, part-of-speech tagging and morphological analysis in general. Our system was based on open-source set of linguistic analysis tools Freeling [Padro et al 2012]. Though in general it showed satisfactory results, we had to fix a number of mistakes, described below. With increase in pre-processing performance, anaphora resolution performance grew accordingly.

The paper is organized as follows. In the section 2 we describe previous work in the field. In the section 3 anaphora resolution machine itself is presented. This section falls into two sub-sections, related to rule-based and machine-learning based modules of the machine. The next section evaluates the performance of the system. We describe typical errors both because of Freeling and because of incompleteness of our algorithm. Then we show how precision and recall measures change with fixing Freeling errors and with various experimental settings. Hybrid approach combining rules and machine learning is presented and its superior performance comparing to 'pure' approaches is shown. In the last section we conclude and propose future work.

2. Related work

Anaphora resolution for English is a well-developed field of natural language processing. First attempt was made in 1964 in an algebra problem-solving system STUDENT ([Bobrow 1964]). Since then the field saw much research.

Typically, anaphora resolution process consists of two steps: first, for each anaphor in the input text, a list of potential antecedent candidates is created. Second, the system decides which of the candidates is the most probable antecedent. Systems can be classified by the way they choose candidates. There are two dimensions of this distinction: types of features for choosing ("restrictional" or "preferential") and methods for choosing using sets of features (traditional rule-based or using machine learning algorithms). "Restrictional" methods are based on discounting candidates which do not satisfy features whereas "preferential" give more preference to those candidates that do satisfy them. An example of restrictional feature is number or gender agreement: candidates which do not agree with anaphoric expression are discarded. Syntax-oriented approaches, for example, Hobbs' algorithm ([Hobbs 1976]), use restrictional approach. An example of preferential feature is centering—giving preference to the most salient (focused) candidate. Detailed though a little outdated overview of anaphora resolution systems for English, features and approaches can be found in [Mitkov 1999]. In a recent evaluation of anaphora resolution for English best system performed with 73.94% F-measure ([Delmonte et al. 2006]).

Anaphora resolution for Russian is not so well-developed as for English. In [Tolpegin et al. 2006] a machine learning approach for third-person pronoun anaphora resolution is presented. Resolution was treated as a classification task, solved using Support Vector Machines. Three types of features were used for classification: distance, positional features and morphological features. The system performed with 62% precision. Unfortunately, recall of the system is unknown.

In [Malkovskiy et al. 2013] another pronominal anaphora resolution system is described. This system used syntactic features along with morphological and distance features and Random forest classification algorithm. Best result was achieved with all features—71% precision. Recall of the system is also unknown.

3. Anaphora resolution system

3.1. Rule-based approach

As stated above, anaphora resolution task falls into two stages: identification of the anaphoric pronoun and identification of its antecedent. First, one has to decide which lexemes are possible anaphoras.

In **An@phora** project we limited ourselves to the following pronouns, loosely separated into three groups: '**personal pronouns**' (*он, она, оно, они, его, её, их, мой*), '**reflexives**' (*себе, свой*), **'relatives'** (*который*). We dropped '*твой*' and '*тот*' from analysis because of their highly discursive nature: the choice of antecedent for these pronouns most of the time heavily depends on deep dialog structure. Moreover, in Russian their antecedent is often only inferred and not expressed by any particular word or multi-word expression. The training set provided by evaluation forum organizers lacks chains with anaphoric "твой" in any form, so it would be difficult to evaluate results even if we decided to handle this pronoun. It should be noted that this is not true for pronoun "мой", which often possesses proper antecedent, especially in the first person narratives.

Antecedent identification within the rule-based module is performed as follows. While reading the given text, we store all the words and noun phrases together with their morphological features. When this stack outgrows a given length (in words), it is shortened from the left to match the threshold. So, this 'analysis window' constantly moves along the text. It allows us to limit antecedent choice to only nearest candidates and not to confuse the system with candidates located far from the anaphoric expression. One can think about 'analysis window' as a kind of shallow salience detector. Experiments showed optimal length of analysis window for Russian texts to be around 23 words; see below.

Upon finding an anaphoric pronoun, the system looks to the left from it in the search of a noun phrase subject to specific constraints. Thus, our anaphora resolution system rule-based module can be classified as simple restrictional one. We presuppose that in most cases the nearest noun phrase abiding to these constraints is the antecedent.

Exact constraints are different for pronoun groups and for some separate pronouns. Simplified example for personal pronoun *она* ('she') will look like: search through all noun phrases within the analysis window. If singular feminine noun phrase in Nominative case[3] denoting animate object is found, consider it to be an antecedent and create an anaphoric chain. If no such noun phrase is found, take the

[3] Our experiments proved that Nominative noun phrases are preferred antecedents for personal pronouns, at least in the provided training set. Introducing this rule increased precision of the anaphora detector.

nearest singular feminine noun phrase and create a chain with it. If no suitable noun phrase is found, consider that the current pronoun is not linked to any antecedent.

Other pronouns have additional peculiarities and constraints. E.g., possessive pronouns of the first person search for their antecedent among first person pronouns, not among noun phrases, reflexives search among both of them, relatives check that there is a comma between anaphoric expression and antecedent, etc.

Despite its generally satisfactory performance (see evaluation section) and unmatched computation speed, rule-based system inherently suffers from its over-simplicity. It is difficult or even impossible to construct all combination of rules manually. Thus, a version of anaphora resolution machine using machine learning approach was designed.

3.2. Machine learning approach

To employ machine learning algorithms we considered anaphora resolution as a classification task: for each anaphoric expression we created a list of candidates and the classifier would label each of them as a possible antecedent or not. We used Random Forest as the main classification algorithm, mainly because it allows ranking features importance. We deliberately used excessive list of features to analyze which make the most contribution in classification. Moreover, Random Forest outperformed most of the popular algorithms (including SVM) for our task in synthetic tests.

As a preprocessing step we performed morphological analysis and simple noun phrase detection. Classifier was implemented using Scikit-learn library ([Pedregosa et al. 2011]). For each anaphoric expression the most probable candidate was returned if its probability was greater than threshold 0.3.

We used the following features for classification:

1. Length of the candidate group in characters
2. Length of the candidate group in words
3. Distance between pronoun and the candidate in characters
4. Distance between pronoun and the candidate in words
5. Distance between pronoun and the candidate in groups
6. Grammatical number of the candidate
7. Grammatical number of the pronoun
8. Do numbers of the candidate and the pronoun agree?
9. Grammatical case of the candidate
10. Grammatical case of the pronoun
11. Do cases of the candidate and the pronoun agree?
12. Is the candidate a proper name?
13. Number of the occurrences of candidate in the text
14. Pronoun type
15. Pronoun itself

Most of these features are fairly standard for this task. Features 12 and 13 are simple salience features. They and distance features are shown to be important (for example, in [Malkovskiy et al. 2013]). Some features, like case agreement, were added without any prior knowledge whether they are helpful or not, to determine their importance.

4. Experiments and evaluation

Evaluation of **An@phora** performance was made against the training set (gold standard), provided by evaluation forum organizers. The set consisted of 92 texts in Russian (69,282 words and 473,445 characters). It contained 2,141 annotated anaphoric chains. As a baseline, we constructed a simple algorithm which links anaphoric expressions to the nearest noun phrase to the left (in the case of reflexives, the nearest pronoun is chosen if it occurs first). This baseline algorithm performed with the following results:

Precision: **0.372**
Recall: **0.357**
F-Measure: **0.364**

4.1. Rule-based system performance

General performance of our system heavily depends on the choice of analysis window length. Our experiments showed that the length of 23 words is optimal with regard to F-measure. After reaching the limit of 23–25 words, analyzer performance quickly degrades, as seen on Fig. 1.

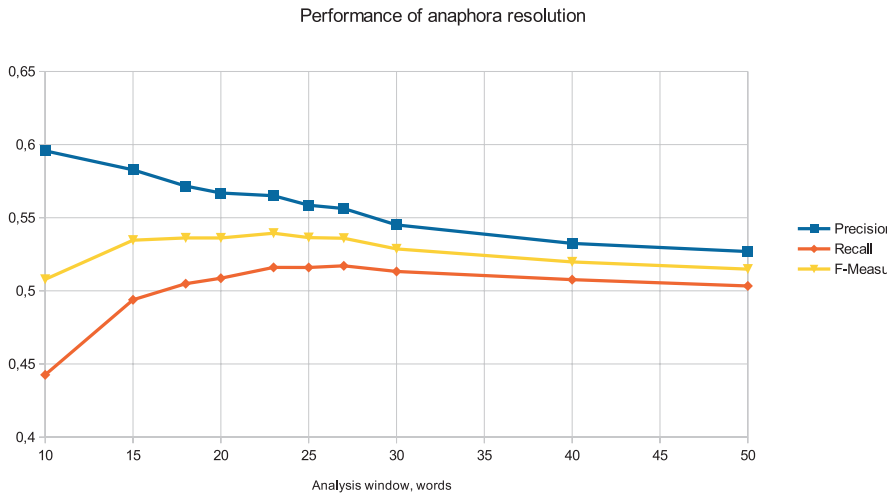

Fig. 1. Performance of anaphora resolution depending on analysis window length

At this optimal length the rule-based module reached **precision 0.565, recall 0.516 and F-measure 0.539** against gold standard, with F-measure 50% higher than the baseline. Attempts to use noun phrases as measure of analysis window length (instead of words) showed worse results with precision only 0.528 and recall 0.494, as well as measuring analysis window in characters (precision 0.47 and recall 0.45 at best). Thus, number of words is a finer setting.

Most frequent errors are related to:

1. Proper names, which are sometimes incorrectly analyzed by Freeling (wrong gender and number).
2. Incorrect choice between inanimate noun in Nominative case and animate noun in some other case. Cf. the expression *'У спутницы Олдмэна на лице траурная вуаль, она несет свои глаза на подносе.'* Our machine links *'она'* to *'траурная вуаль'*, as it is the nearest noun phrase, and additionally Nominative. *'Спутницы'* would win the contest only if it were Nominative, but it is not. Experiments with lending animate nouns bonuses independent of their case degraded performance.
3. Antecedent beyond the limits of analysis window. It is mostly found in encyclopedic articles with a lot of impersonal sentences, whose real subject is the article headword, and affects pronouns like *'себя'* and *'собой'*.
4. Cases of cataphora. Our system does not detect it, and finds incorrect antecedents to the left of the anaphoric expressions, when they are in fact on the right.
5. Direct speech (cases when our system links anaphoric expression to a noun group inside quotes). As in the above mentioned case with pronoun *'твой'*, proper handling of such cases demands processing of dialog structure, and we consider this to be future work.

4.2. Machine learning system performance

Evaluation of machine learning (ML) approach was performed on two random subcorpora as test sets while the rest of gold standard was used as a training set. This was done in order not to over-fit classifier using the same data for training and testing. Size of subcorpora are presented in the table 1.

Table 1. Size of subcorpora for ML system evaluation

	Test set, texts	Train set, texts
Subcorpus 1 (further S1)	5	106
Subcorpus 2 (further S2)	13	95

F-measure of machine learning (ML) resolution is generally worse than for rule-based (RB) system, however, precision is consistently higher. The results are presented in the table 2.

Table 2. Comparison of ML and RB systems performance on subcorpus S1 and subcorpus S2

	Precision	Recall	F-measure
Rule-based, S1	45.02%	46.58%	45.79%
Machine Learning, S1	52.54%	43.51%	47.60%
Rule-based, S2	63.57%	56.45%	59.80%
Machine Learning, S2	65.11%	45.67%	53.69%

Analysis of feature importances in both models shows this order for features with importance > 0.05:

1. Distance in characters
2. Distance in words
3. Distance in groups
4. Length of candidate in characters
5. Pronoun
6. Number of the occurrences of candidate in text
7. Case of the candidate
8. Type of the pronoun

As we can see, distance appears to be the most important feature, whereas number and case agreement are much less important. Interestingly, distance in characters appears to be more important than distance in words (0.193 and 0.121, respectively). This needs further analysis because this result is far from obvious.

4.3. Influence of morphologic processing on system performance

It turns out that performance of preliminary NLP steps, such as morphological analysis, has crucial influence on performance of anaphora resolution. E.g., we discovered several inconsistencies in Freeling handling of Russian pronouns (supposedly, not Freeling itself should be blamed for that, but the corpus on which its Russian module had been trained). Fortunately, Freeling is very flexible and allows to fine-tune its morphologic analysis model. Among others, we had to fix number and case probabilities for *'его'* and add missing gender annotation to almost all personal pronouns: more than fifteen corrections total.

We compared performance of our rule-based anaphora resolution machine with original and fixed Freeling. The results are given in the table 3.

Table 3. Performance increases after fixes to morphology processing of anaphoric pronouns

	Original Freeling	**Fixed Freeling**
Precision	0.493	0.565
Recall	0.424	0.516
F-Measure	0.456	0.539

This dramatic difference comes as no surprise. We extensively use number, gender and case features of anaphoric expressions to check their agreement with antecedents. Thus, insufficient or outright incorrect morphological processing directly influences anaphora resolution performance.

It should be noted that of course Freeling errors are not limited to pronouns. Occasionally it wrongly detects noun case or gets stuck on nominalized adjectives. We experimented with a text from gold standard 1,389 words long. Our machine

detected anaphoric chains in this text with precision 0.3, recall 0.28 and f-measure 0.29. However, after we manually fixed Freeling output for all words of the text, precision raised to 0.31, recall to 0.3 and f-measure to 0.3. Among others, manual post-processing fixed treating proper surname *'Одиноков'* as genitive plural and let pronoun *'их'* link to correct antecedent instead of this.

At the same time, all in all we had to make only 86 corrections to the annotation of the text containing 1,389 words. Thus, only 6% of Freeling output demanded any manual intervention (and a lot of them only slight one, like adding animation property). We consider it a sufficient degree of quality. However, we also plan to return all our corrections to Freeling developers to be incorporated in the next release.

4.4. Hybrid approach

Machine learning approach shows higher precision than rule-based approach but lower recall, thus we created a hybrid system to improve both results using advantages of each method. Low recall with high precision means that if the system returns a result it is confident about it. So we integrated a new stage in our rule-based pipeline: for each pronoun we tried to predict antecedent with ML. If it couldn't predict, we used rule-based approach. This method showed improvement in average: when rule-based approach shows low F-measure, hybrid one improves the result drastically, when the rule-based approach shows high F-measure, hybrid one may lower overall results but not critically. See table 4 for comparison.

Table 4. Comparison of ML, RB and hybrid systems' performance on subcorpora S1 and S2

	Precision	**Recall**	**F-measure**
Rule-based, S1	45.02%	46.58%	45.79%
Machine Learning, S1	**52.54%**	43.51%	47.60%
Hybrid, S1	49.49%	**53.72%**	**51.52%**
Rule-based, S2	63.57%	**56.45%**	**59.80%**
Machine Learning, S2	**65.11%**	45.67%	53.69%
Hybrid, S2	62.22%	**56.46%**	**59.20%**

Thus, hybrid approach seriously outperforms both rule-based and machine learning ones on S1 subcorpus and is almost on a par with rule-based algorithm on S2 subcorpus.

5. Conclusion and future work

We presented **An@phor**a—a system for automated anaphora resolution in Russian texts. It is freely available under open-source GPL license and can be tested through online demo.

Our approach includes using Freeling as an underlying morphologic analysis layer and a combination of rules and machine learning model to ensure better anaphora resolution quality. Our separate rule-based module, tested against training set provided by evaluation forum organizers, showed **F-measure of 0.539 with precision 0.565 and recall 0.516.** General hybrid module, tested on two different subcorpora from the training set, showed **F-measure from 0.51 to 0.59**. General improvement of F-measure in comparison to a simple baseline algorithm is 40% to 62%.

At the same time, there is still room for future improvement. We plan to train our classifier on a larger corpus, as 70 thousand words from gold standard is clearly not enough. Rule-based module can also be improved to consider direct speech issues, cataphora and sentences interaction. Finally, handling of "твой" and "тот" anaphoric expressions should be implemented.

6. Acknowledgments

The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2014.

References

1. *Delmonte, R., Bristot, A., Piccolino Boniforti, M. A., and Tonelli, S.* (2006). Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach, In Proc. of ROMAND 2006, Trento, pp. 3–10.
2. *Hobbs, Jerry R.* (1976) Pronoun Resolution. Research Report 76-1, Department of Computer Sciences, City College, City University of New York. August 1976.
3. *Lluís Padró and Evgeny Stanilovsky* (2012), FreeLing 3.0: Towards Wider Multilinguality, Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey.
4. *Mitkov R.* (1999) Anaphora resolution: the state of the art.—School of Languages and European Studies, University of Wolverhampton
5. *Pedregosa* et al. (2011) Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825–2830, 2011.
6. *Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii* (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. Proceedings of the Demonstrations Session at EACL 2012.
7. *Tolpegin P. V., Vetrov D. P., Kropotov D. A.* (2006) Algorithm for machine-learning based automated resolution of third person pronominal anaphora [Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения]. Proceedings of Dialog-2006 conference (Moscow, Bekasovo, 2006)
8. *Malkovsky M. G., Starostin A. S., Shilov I. A.* (2013) A method for pronominal anaphora resolution in the course of syntactic analysis [Метод разрешения местоименной анафоры в процессе синтаксического анализа] Proceedings of Sworld conference, pp. 41–49.