The Implementation of the Mention-Ranking Approach to Coreference Resolution in Russian

Anna Kupriianova
annkupriyanova26@gmail.com

Ivan Shilin
shilinivan@corp.ifmo.ru

Gerhard Wohlgenannt
wolkinger@gmail.com

Liubov Kovriguina
lyukovriguina@corp.ifmo.ru

ITMO University
Saint-Petersburg, Russian Federation

CEUR-WS.org/Vol-2233/Paper_15.pdf

Abstract

Coreference resolution is a fundamental ingredient for many downstream tasks in natural language-based applications. For Russian language, the work in coreference resolution is very limited. In this publication, we present a system inspired by the mention-ranking approach, which improves the state-of-the-art F1 score from 0.63 to 0.71, measured with the B^3 metric. We evaluate various sets of feature combinations, and also discuss the limitations of the presented work.

Keywords: *coreference resolution in Russian, mention-pair model, neural network based coreference system, RuCor, FastText model.*

1 Introduction

Coreference resolution is an important problem in many natural language processing tasks. It can support, i.e., automatic text summarization, knowledge extraction, objects' identification in dialogue and translation systems [Toldova, Ionov 2017]. The term *coreference* denotes the relation between the parts of a text (mentions), that refer to the same real world entities, for example the mention of a person name in a text, and of a pronoun that refers to this same person. Thus, the task of coreference resolution is to find and group all the mentions in the text according to their referents. Mentions are typically represented by noun phrases (NPs), named entities and pronouns, except the cases of abstract anaphora, where the anaphoric pronoun refers to the whole preceding sentence and not to noun phrases (NPs) or pronouns [Nedoluzhko, Lapshinova-Koltunski 2016]. In the pair of mentions the first one (full mention) is called the antecedent while the second one is an anafor. In the broader field of coreference resolution, there are two main tasks: mention extraction and coreference resolution in narrower sense (mention clustering) [Sysoev et al. 2017]. Mention extraction finds textual expressions which are possible elements of coreference chains in unstructured data, whereas coreference resolution groups

mentions into clusters, which refer to a single real-world entity. The system presented here focuses on the second task, i.e. coreference resolution in the narrower sense.

For Russian language, the work on coreference resolution is limited. Results reported on the RuCor[1] [Toldova et al. 2014] coreference corpus are 60.48 of F1 score for the B^3 metric by Toldova and Ionov [Toldova, Ionov 2017], and 63.12 F1 by Sysoev et al. [Sysoev et al. 2017]. The mentioned work applies rule-based and "classical" machine-learning methods like decision trees or logistic regression. In this work, we present an approach using neural networks based on an adapted version of the mention-ranking model by Clark and Manning [Clark, Manning 2016]. With this architecture, we manage to outperform previous work with an achieved F1-score of 0.7131. The B^3 metric [Bagga, Baldwin 1998, Amigo et al. 2009] is a clustering metric, which evaluates a gold-standard clustering of mentions against a system-produced clustering.

The paper is structured as follows: After an overview of related work in Section 2, Section 3 introduces the system architecture, and Section 4 discusses the features used, and how features are combined into three different sets. The following section (Section 5) provides the evaluation details for those feature sets and compares the results to the state-of-the-art. Furthermore, difficult cases are discussed. The paper concludes with Section 6.

2 Related Work

Existing approaches to coreference resolution can be divided into heuristic [Hobbs 1978, Boyarski et al. 2013] and based on machine learning (ML) algorithms [Rahman, Ng 2009, Ng 2008, Clark, Manning 2016]. Heuristic methods are built upon a handmade set of rules, which is time- and labour-consuming to construct. On the contrary, ML-based approaches are faster and easier to develop, but they depend on the availability of a coreference dataset of sufficient size and quality to apply supervised learning methods.

Recent advances in coreference resolution for English language include the work of Clark and Manning [Clark, Manning 2016] on a cluster-ranking algorithm that handles entity-level information and eliminates the disadvantages of the mention-pair models. The main benefit of this neural network-based method is that it can distinguish beneficial cluster merges from harmful ones. Central parts of the system architecture used for this publication are inspired by this work.

In general, the state-of-the-art results for Russian language are lower than for English, work on Russian language is rather limited so far [Sysoev et al. 2017]. For Russian language, coreference resolution became a more active research topic with the release of a tagged coreference corpus (RuCor) in 2014 [Toldova et al. 2014]. Toldova and Ionov [Toldova, Ionov 2017] compare rule-based and ML-based methods for coreference resolution using this RuCor dataset, with slightly better performance for the ML-based methods. They reach 31.56 for predicted mentions and 60.48 for gold mentions with the B^3 metric for the RuCor corpus. *Predicted mentions* refers to coreference resolution for mentions which were automatically extracted with a mention extraction module. Sysoev et al. [Sysoev et al. 2017] tackle both mention extraction and coreference resolution. For mention extraction, they use a number of linguistic, structural, etc., features and apply classifiers such as logistic regression, Jaccard Item Set mining and random forest, and

[1] http://rucoref.maimbava.net/

reach an F1 of 63.12 for the gold mentions from the RuCor dataset. In comparison, our approach provides an F1-score about 8 points above previous work.

3 System Architecture

We present a coreference resolution system based on the mention-ranking model by Clark and Manning [Clark, Manning 2016]. The core of the system is a feedforward neural network. Its topology is shown in Figure 1. In a nutshell, the network can be divided into two parts: a mention-pair encoder and a mention-ranking model. The implementation of the system is available on github[2]. The source code was written in Python, using Keras and Tensorflow for the neural network models.

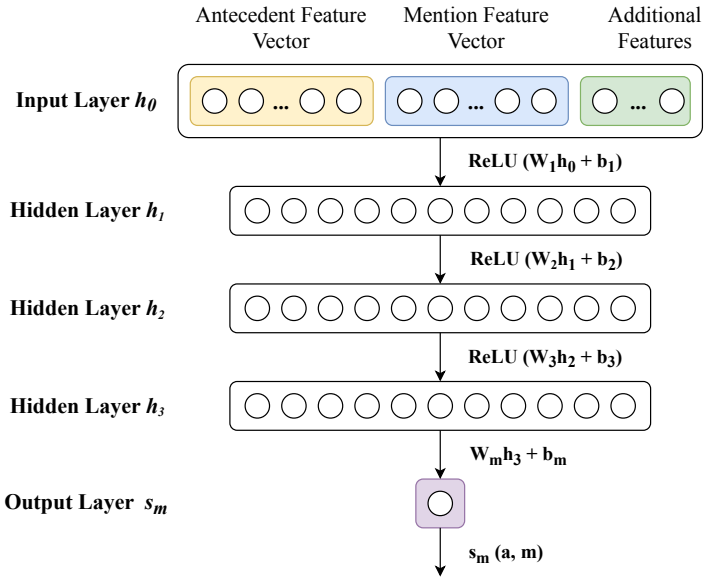

Figure 1: Neural Network Topology

3.1 Mention-Pair Encoder

The purpose of the mention-pair encoder is to transform a pair of a mention m and its potential antecedent a into their distributed representations. The mention-pair encoder is implemented as a feedforward neural network with three fully-connected hidden layers of rectified linear units (ReLU):

$$h_i(a, m) = max(0, W_i h_{i-1}(a, m) + b_i) \quad (1)$$

where $h_i(a, m)$ is an output of the i-th hidden layer for a pair of mention m and its potential antecedent a. W_i is a weight matrix and b_i is the bias for the i-th hidden layer.

The input layer of the mention-pair encoder takes a vector of features of a mention and its potential antecedent as well as additional pair features (all the features are described in Section 4). The output of the last hidden layer is the distributed representation of the pair which is used as an input to the mention-ranking model.

[2] https://github.com/annkupriyanova/Coreference-Resolution

3.2 Mention-Ranking Model

The purpose of the mention-ranking model is to estimate the score of coreference compatibility for the pair of a mention m and its potential antecedent a. To compute this score one applies one fully-connected layer to the distributed representation of the pair $r_m(a, m)$,

$$s_m(a, m) = W_m r_m(a, m) + b_m \tag{2}$$

where $s_m(a, m)$ denotes a score of coreference compatibility of a pair of mention m and its potential antecedent a, $r_m(a, m)$ is the distributed representation of this pair.

3.3 Training objective

For pretraining, which determines the initial configuration of the model parameters, we used the following objective function,

$$-\sum_{i=1}^{N}\Big[\sum_{t\in T(m_i)}\log p(t, m_i) + \sum_{f\in F(m_i)}\log(1 - p(t, m_i))\Big] \tag{3}$$

where $T(m_i)$ and $F(m_i)$ are sets of true and false antecedents of a mention m_i respectively, and $p(a, m_i) = sigmoid(s(a, m_i))$.

The main training objective is a slack-rescaled max-margin which penalizes different types of errors:

$$\sum_{i=1}^{N}\max_{a\in A(m_i)}\Delta(a, m_i)(1 + s_m(a, m_i) - s_m(\hat{t_i}, m_i)) \tag{4}$$

where $A(m_i)$ is a set of candidate antecedents of a mention m_i, $\hat{t_i}$ is a highest scoring true antecedent of mention m_i:

$$\hat{t_i} = \arg\max_{t\in T(m_i)} s_m(t, m_i) \tag{5}$$

and $\Delta(a, m_i)$ is a cost function for different types of mistakes:

$$\Delta(a, m_i) = \begin{cases} \alpha_{FN} & if\ a = NA \wedge T(m_i) \neq NA \\ \alpha_{FA} & if\ a \neq NA \wedge T(m_i) = NA \\ \alpha_{WL} & if\ a \neq NA \wedge a \notin T(m_i) \\ 0 & if\ a \in T(m_i) \end{cases} \tag{6}$$

where FN stands for False New mistake, FA for False Anaphor, and WL for Wrong Link.

4 Feature Sets and Models

For creating the coreference resolution models, we designed three feature sets. We will compare the results for the individual models in the evaluation section. Table 1 shows how the features are partitioned into our feature sets.

The list of features is inspired by previous work on English and Russian coreference resolution. The feature set I is a reduced version of the features used by Clark and Manning [Clark, Manning 2016]. In feature set II we removed some of the word embedding features, and added features of explicit indication of morphological characteristics and

Table 1: Feature Sets

Feature	**Model I**	**Model II**	**Model III**
Word embedding of the head word of the mention	+	+	-
Average word embedding of all words in the mention	+	-	-
Cosine similarity between the vectors of mentions heads	-	-	+
Gender	+	+	+
Number	+	+	+
Animacy	+	+	+
Exact string match	+	+	+
Head string match	+	+	+
Partial string match	+	+	+
Distance between the mentions in intervening mentions	+	+	+
Distance between the mentions in sentences	-	+	+
Gender match	-	+	+
Number match	-	+	+
Animacy match	-	+	+
Both mentions are proper nouns	-	+	+
Both mentions are pronouns	-	+	+
Antecedent is a pronoun	-	+	+
Anafor is a pronoun	-	+	+

POS-tag agreement, which is highly relevant for Russian as a morphologically rich language[3]. And in feature set III the cosine similarity between the vectors of mentions heads completely replaces the word embeddings.

For each mention, we build a feature vector. A mention consisting of one word, is represented by a single vector (word embedding). If the mention consists of a group of words, is represented by the average of the embedding vectors of each word in the group. We use word embeddings pre-trained with FastText on the Wikipedia corpus[4].

5 Experiments and Results

This section describes the experimental setup, especially the dataset, and provides the results of the evaluations for the three models introduced in Section 4 in comparison with existing work. Finally, we discuss some of the difficulties and limitations observed with the current architecture.

[3]Morphological annotation, lemmatization and word embedding models were borrowed from[Kovriguina et al. 2017]

[4]https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

5.1 Experimental Setup

With regards to the *dataset*, we used the RuCor dataset to train and test the coreference resolution system. RuCor is the first open corpus with coreference annotations for the Russian language. It comprises short texts and text fragments of different genres: news, fiction, scientific papers, etc. All the texts are tokenized, split into sentences and parsed syntactically and morphologically. In the corpus, mentions are limited to NPs that refer to real-world entities. Thus, abstract and generic NPs as well as bridging relations and coreference relations with a split antecedent are not annotated. The RuCor dataset contains 181 texts with 156637 tokens and 3638 coreference chains. Based on these numbers, in model training we used a 60-20-20 percent split to create the training, validation, and test datasets.

In terms of the settings of the neural network models, we minimized the training objectives with the Adam and RMSProp optimizers. For regularization, we applied dropout with a rate of 0.3 on the output of each hidden layer.

5.2 Experiments

For the feature sets three models of the neural network were developed. All the models were pretrained. Pretraining is performed for an initial setup of the weights, and uses binary cross-entropy to compute the loss when training only the mention-ranking model. The results of the experiments are shown in Table 2.

Table 2: Experiment Results

	AUC	B^3 metric		
		Precision	Recall	F1 score
Training				
Model I	0.6538	0.6397	0.6711	0.6550
Model II	**0.8280**	**0.7170**	**0.7092**	**0.7131**
Model III	0.7870	0.6783	0.6902	0.6842

According to Table 2 Model II shows the best results. We suppose that word embeddings and explicit indication of the matches in morphological attributes and POS-tags positively influence the results. In Model III word embeddings are replaced with cosine similarity between the embeddings for the mentions' head words. It appears that is may not be sufficient for the identification of their semantic similarity. Model I has the lowest score which might be explained by the absence of the explicit indication of matches in features, which, as stated above, lowers the model quality. Moreover, the non-proportional lengths of the vectors for different features, with 300-dimensional vectors for the head word of each mention in a pair and all the words of these mentions, in contrast with only 38-dimensional vector for the other features, might influence the outcome.

5.3 Comparison

In Table 3 we compare the results of our system with the state-of-the-art open coreference resolution systems for Russian by Tolodova and Ionov [Toldova, Ionov 2017] and Sysoev

et al. [Sysoev et al. 2017]. We compare the B^3 metric on the gold mentions, i.e. mentions from the RuCor corpus (not for mentions automatically extracted from the text). All our models surpass existing work with regards to F1 score, with model II giving the best results.

Table 3: Comparison

Model	B^3 metric		
	Precision	Recall	F1 score
Model I	0.6397	0.6711	0.6550
Model II	0.7170	**0.7092**	**0.7131**
Model III	0.6783	0.6902	0.6842
Toldova and Ionov [Toldova, Ionov 2017]: MLUpdated	**0.7937**	0.4860	0.6029
Toldova and Ionov [Toldova, Ionov 2017]: NamedEntities	**0.7937**	0.4886	0.6048
Toldova and Ionov [Toldova, Ionov 2017]: Word2vec	0.7925	0.4864	0.6028
Sysoev et al. [Sysoev et al. 2017]: log. regr. + Jaccard Item Set mining	0.6014	0.6103	0.6055
Sysoev et al. [Sysoev et al. 2017]: random forest	0.7389	0.5516	0.6312

The comparison baselines can be briefly described as follows (for details see Tolodova and Ionov [Toldova, Ionov 2017], and Sysoev et al. [Sysoev et al. 2017]): The MLUpdated model implements a ML-based decision tree classifier. The NamedEntities model takes into account semantic information in the form of the lists of possible named entities. This allows it to compare the mentions' semantic classes. The Word2vec model uses word embeddings for evaluating the semantic compatibility of the mentions' heads (we used the same feature in our Model I and Model II). Sysoev et al. [Sysoev et al. 2017] use a common set of features, which is fed into various classifiers such as logistic regression with Jaccard Item set mining, or a random forest.

5.4 Error Analysis and Results Discussion

Here, we outline some of the problems and errors which have been discovered in the analysis of the predictions of the neural network:

1. Some errors are caused by the wrong annotations in the RuCor coreference corpus, esp. wrong lemmas or morphological attributes in the corpus. For example, for the word "*дотком*" (dotcom) two different lemmas were found – "*доткома*" and "*доткомом*".

2. Direct speech mistakes: Pronouns "*я*" (I) and "*ты*" (you), if used in a dialogue by different speakers, can be coreferential in a certain context. For example, in case of the following dialogue:

– "*Я сегодня выполнил работу за два дня.*" (Today I have done the work for two days.)

– "*Ты - молодец!*" (You did well!)

The pronouns "*я*" (I) and "*ты*" (you) have the same referent – the first speaker. However, the neural network makes this kind of mistakes because there is no information about speakers in the dataset.

3. Context mistakes: They arise when the coreference relation gets evident only after the analysis of the mentions context. For example, the coreference relation between the mentions "*Их Сиятельство*" (Their Majesty / Highness) and "*женщина в черном капоте*" (the woman wearing black dressing gown) is not clear without the analysis of the context parts of the text. Such types of mistakes can be explained with the difficulty of formalizing semantic information. One possible solutions for this problem might be the use of word embeddings for a longer context or even for the whole text.

4. Split anafor mistakes: For example, the mentions "*они*" (they) and "*Иван Тихонович и Татьяна Финогеновна*" (Ivan Tihonovich and Tatyana Finogenovna) are coreferential. But the network makes the mistake because in the dataset the morphological attributes (gender, number, animacy, etc.) are identified only for the head elements of the mentions. And in the above stated case there are two heads in the second mention and they differ in gender with each other and in number with the head of the first mention.

6 Conclusion

We have presented a coreference resolution system implementing the mention-ranking approach, and experimented with different sets of feature combinations and evaluated their impact on system quality with the B^3 metric. Our best model provides an F1 score of 0.7131 on the RuCor dataset, and exceeds existing work by around 8 points. In future work, there are a number of directions to improve model quality: (i) experimenting with other network architectures, (ii) hyperparameter tuning, and (iii) adding more relevant features into the training process, and finally (iv) applying a cluster-ranking approach to capture additional entity-level information. Furthermore, we plan to extend our system with a mention extraction module.

Acknowledgments

L.Kovriguina acknowledges support from the Russian Fund of Basic Research (RFBR), Grant No. 16-36-60055. Furthermore, the work is supported by the Government of the Russian Federation (Grant 074-U01) through the ITMO Fellowship and Professorship Program.

References

[Hobbs 1978] Hobbs. J. R (1978) Resolving pronoun references. *//Lingua – Volume 44*, p. 311–338, 1978.

[Boyarski et al. 2013] Boyarski K.K., Kanevski E.A., Stepukova A.V. (2013) Viyavlenie anaforicheskih otnosheni pri avtomaticheskom analize teksta [Identification of anaphoric relations in automatic text analysis]. *//Nauchno-tehnicheski vestnik informacionnyh tehnologi, mehaniki i optiki [Scientific and Technical Herald of Information Technologies, Mechanics and Optics]*, s. 108–112, 2013. (In Russian) = Боярский К. К., Каневский Е. А., Степукова А. В. Выявление анафорических отношений при автоматическом анализе текста. *//Научно-технический вестник информационных технологий, механики и оптики*, с. 108–112, 2013.

[Rahman, Ng 2009] Rahman A., Ng V. (2009) Supervised models for coreference resolution. *//Proceedings of the 2009 conference on empirical methods in natural language processing*, p. 968–77. Singapore, 2009.

[Ng 2008] Ng V. (2008) Unsupervised models for coreference resolution. *//Proceedings of the 2008 conference on empirical methods in natural language processing*, p. 640–9. Honolulu, 2008.

[Clark, Manning 2016] Clark K., Manning C. D. (2016) Improving Coreference Resolution by Learning Entity-Level Distributed Representations. *//Association for Computational Linguistics Proceedings – Volume 1*, p. 643–653. Berlin, 2016.

[Toldova et al. 2014] Toldova S. Ju., Roytberg A., Nedoluzhko A., Kurzukov M., Ladygina A., Vasilyeva M., Azerkovich I., Grishina Y., Sim G., Ivanova A., Gorshkov D. (2014) Evaluating Anaphora and Coreference Resolution for Russian. *//Computational Linguistics and Intellectual Technologies: "DIALOG 2014"*, p. 681–695. – M.: RGGU, 2014.

[Toldova, Ionov 2017] Toldova S., Ionov M. (2017) Coreference Resolution for Russian: The Impact of Semantic Features. *//Computational Linguistics and Intellectual Technologies. International Conference "Dialog 2017" Proceedings*, p. 339–349. – M.: M., 2017.

[Sysoev et al. 2017] Sysoev A., Andrianov I., Khadzhiiskaia A. (2017) Coreference Resolution in Russian: State-of-the-Art Approaches Application and Evolvement. *//Computational Linguistics and Intellectual Technologies. International Conference "Dialog 2017" Proceedings*, 16(23):327–347.

[Nedoluzhko, Lapshinova-Koltunski 2016] Nedoluzhko A., Lapshinova-Koltunski E. (2016) Abstract Coreference in a Multilingual Perspective: a View on Czech and German. *//Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, p.47–52.

[Bagga, Baldwin 1998] Bagga A., Baldwin B. (1998) Entity-based Cross-document Coreferencing Using the Vector Space Model *//Proceedings of ACL'98, 1998, Montreal, Quebec, Canada*, p.79–85.

[Amigo et al. 2009] Amigo E., Gonzalo J., Artiles J., Verdejo F. (2009) A comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints *//Information Retrieval – Volume 12*, p. 461–486, 2009.

[Kovriguina et al. 2017] Kovriguina, L., Shilin, I., Putintseva, A., Shipilo, A. (2017) Russian Tagging and Dependency Parsing Models for Stanford CoreNLP Natural Language Toolkit. *//International Conference on Knowledge Engineering and the Semantic Web*, p.101–111. – Springer, 2017.