Using Semantic Information for Coreference Resolution with Neural Networks in Russian

Ilya Azerkovich(✉)

Higher School of Economics, Moscow, Russia
ilazerkovich@edu.hse.ru

Abstract. This paper describes an experiment aimed at improving the quality of coreference resolution for Russian by combining one of the most recent developments in the field, employment of neural networks, with benefits of using semantic information. The task of coreference resolution has been the target of intensive research, and the interest at using neural networks, successfully tested in other tasks of natural language processing, has been gradually growing. The role that semantic information plays for the task of coreference resolution has been recognized by researchers, but the impact of semantic features on the performance of neural networks has not been yet described in detail. Here we describe the process of integrating features derived from open-source semantic information into the coreference resolution model based on a neural network, and evaluate its performance in comparison with the base model. The obtained results demonstrate quality on par with state-of-the-art systems, which serves to re-establish the importance of semantic features in coreference resolution, as well as the applicability of neural networks for the task.

Keywords: Natural language processing · Coreference resolution · Neural networks · Semantic relatedness · Russian language

1 Introduction

Coreference resolution, as an important step at machine translation, information extraction, text summarization, etc., is among the most relevant tasks of natural language processing. Two expressions can be considered coreferent if they refer to one and the same real-world entity. Consequently, the goal of automated coreference resolution is to extract chains of mentions, referring to the same entity, from the text.

The algorithms of automated coreference resolution have been created since the middle of XXth century. At first these algorithms have been mostly empiric, based on the rules suggested by its developer, such as the classic algorithm described in [1]. Later the work actively began on the family of algorithms based on machine learning methods and using big data for training [2] or [3], but the rule-based algorithms are also successfully used, for example in the Stanford coreference parser [4]. Recently, followed by the rise of interest towards neural networks, the works aimed at using them for various NLP tasks, including coreference resolution, have started to appear [5, 6].

© Springer Nature Switzerland AG 2020
W. M. P. van der Aalst et al. (Eds.): AIST 2019, CCIS 1086, pp. 85–93, 2020.
https://doi.org/10.1007/978-3-030-39575-9_9

Main types of features used in algorithms of automated coreference resolution include morphological, syntactic, string-based and distance ones. Semantic information, if used, usually is presented as information about named entities or compatibility in terms of top-level ontology nodes (e.g. in [2] or [7]). Features, derived by more detailed analysis, such as semantic relatedness measures, are seldom used despite their potential effectiveness shown in a number of works, such as [8].

Development of algorithms of automated coreference resolution for Russian began later than for English. This may partly be due to algorithms relying upon resources in the corresponding language, which for Russian exist on a much smaller scale. Research, describing systems of coreference resolution in Russian, does exist (see e.g. competition results of [9]), and attempts at using semantic information for analysis are also being made. This paper describes integrating semantic relatedness measures, calculated from open-source data, into a neural network-based algorithm, oriented at Russian language. The achieved results suggest that, while using neural networks for the task of coreference resolution in Russian could be more effective than other methods, by using semantic features further improvement could be achieved.

2 Related Work

Machine learning-based algorithms are the most actively developed class of methods for automated coreference resolution. They can be grouped into several general classes based on structure, the main of them being: the mention-pair models, suggested in the seminal work of Soon et al. [2]; the entity-mention models, described e.g. in [10], which introduces mention clustering and cluster-based features; ranking models, which consider several candidate mentions (e.g. [11]) or mention clusters [12] as possible antecedents. The usage of perceptrons and neural networks has been researched, among others, for a mention ranking model in [13], and for a cluster ranking model in [6].

Russian language-oriented research has begun to develop actively relatively recently, with a breakthrough becoming possible due to the publication of the RuCor corpus, used for the RuEval-2014 competition of coreference resolvers [9]. While most participants of the competition employed different pair-based models for the analysis, clustering approach has been also adopted in [14].

Using semantic information for the task of coreference resolution in English has been studied in several papers. Particularly, the usage of semantic similarity measures as features has been researched in [8] and [15]. For Russian language, the implementation of semantics-based features in the form of hypernym chains and gazetteers has been described in [16], and an attempt at using the data of Wikipedia articles was made in [17].

In this paper we use the semantic information, obtained from the Russian thesaurus RuThes-lite as well as from Wikipedia, to calculate semantic relatedness measures to be used as features in the machine learning algorithm. We attempt to realize a mention-ranking algorithm using neural networks for predicting coreference, based on the one described in [6].

3 System Architecture

In this work we describe a mention-ranking model, derived from the algorithm introduced in [6]. It is based on a feedforward neural network, consisting of two main modules: the mention-pair encoder and the mention-ranking layer. The model was developed, relying on the existing open-source solutions with the use of Keras and Tensorflow libraries for the Python programming language.

3.1 Mention-Pair Encoder

The first module of the network was tasked with transforming the input to its distributed representation. The model receives as input the vector, consisting of embeddings of an antecedent and its potential anaphor and their features as well as several additional pair features (sets of features are described in detail in Sect. 4), and its output is then fed to the mention-ranking model.

Structurally the encoder presents a three-layer fully connected neural network with hidden layers of rectified linear units (ReLU):

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

Here $h_i(a, m)$ is the output of the i-th layer with the input of mention m and its potential antecedent a, Wi is a weight matrix, and b_i is the layer's bias.

3.2 Mention Ranking Model

The second module in the network, the mention-ranking model, estimates the coreference score of the pair of a mention m and its possible antecedent a. As the input it accepts the distributed representation of the pair, the output of the mention-pair encoder. It is represented by a single fully connected layer with the sigmoid activation function:

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

Here s_m is the coreference score of the pair, and r_m is its distributed representation.

3.3 Training the Network

Pretraining the neural network has been determined by [6] among others as an important step in its development. For the pretraining of our network the following function was used:

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

$T(m_i)$ is the set of all true antecedents of the i-th mention m_i, $F(m_i)$ is the set of all false antecedents of the same mention, and $p(t, m_i) = sigmoid(t, m_i)$.

As the training objective, the slack-rescaled max-margin was used. First, the highest-scoring antecedent of the mention m_i was found:

$$\hat{t}_i = \text{argmax}_{t \in T(m_i)} s_m(t, m_i) \tag{4}$$

Then, the loss function was calculated:

$$\sum_{i=1}^{N} \max_{a \in A(m_i)} \Delta(a, m_i)\left(1 + s_m(a, m_i) - s_m\left(\hat{t}_i, m_i\right)\right) \tag{5}$$

$A(m_i)$ here is the set of all possible antecedents of the mention m_i, and $\Delta(a, m_i)$ is the mistake-specific cost function:

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

Here α_{FN}, α_{FA} and α_{WL} denote costs for different error types: "false new", "false anaphoric" and "wrong link", correspondingly. The values used were {0.5, 0.5, 1}.

For training the model, the Adam optimizer was used. The dropout rate was set to 0.3 for all hidden layers.

4 Feature Sets

For the purposes of this research, the performance of different models with two different feature sets was compared. The default set consisted of string-based, morphological, lexical and distance features, generally used in coreference resolution algorithms. The second feature set also included as features measures of semantic relatedness, calculated from semantic information from two external sources: the Russian Wikipedia and a Russian thesaurus RuThes-Lite.

4.1 Default Model

The default feature set consisted of features, traditionally used for the task of coreference resolution ([2, 6, 7], among others). It combined separate morphologic and lexical features of the mention and the antecedent with features defined for the pair, such as distance between members and matches in strings or POS-tags. As the lexical features of the mentions, the word embeddings were used. The embeddings were obtained from Wikipedia corpus using FastText. If a member of the pair was a noun phrase, the representation of its head was used. An attempt to use the average word embedding of all words in a phrase as well was made, but it yielded worse results.

The complete list of features is given in Table 1 below:

Table 1. The default feature set

Feature class	Features
String-based	Full string match Head string match Partial string match
Distance	Number of NPs between members
Morphological	Number Gender Animacy Number match Gender match Animacy match Both members are proper One of members is a pronoun Both members are pronouns
Lexical	Word embeddings of the NP head

4.2 Semantic Information Extraction

The alternative feature set we used in our research was enriched with measures of semantic relatedness between members of mention-antecedent pairs. To generate these features, the semantic information from two publicly available sources was analyzed. One of them is RuThes-Lite: a thesaurus of Russian, including 55 000 entities that correspond to 158 000 lexical entries [18]. The structure of RuThes-Lite is similar to that of WordNet, with concepts in the thesaurus linked to each other by the set of labeled relations, including IS-A, PART-WHOLE and a number of associative relations. The other source was the Russian segment of Wikipedia. While being smaller than the English one (~1.5 mln articles, compared to ~5 mln articles), it is still one of its largest, making it an important knowledge source. The reason Wikipedia was chosen as a source is its category structure, which can be analyzed in similar terms to a thesaurus: each Wikipedia article is placed within one or several categories that, in their own turn, can be categorized further.

This allowed to analyze the category structure of Wikipedia as a graph, in the same way as Ruthes was analyzed. Categories were considered as graph nodes, and relations of inclusion between them – as edges. Articles belonging to a category were considered terminal nodes of the graph. This representation also made it possible to apply the same semantic relatedness metrics to both of our selected sources.

The following set of measures of semantic relatedness was used for analysis: the path-based measures, suggested by [19, 20] and [21], and information content-based measure, suggested by [22]. For each pair of mention and antecedent the values of the

metrics between the head lemmas of both groups were calculated. If any of them was ambiguous, for the combinations of possible meanings the average and maximum values of the metric were calculated and used as features.

5 Experiment Setup and Evaluation

5.1 Corpus Data

The model was trained and tested on the data of RuCor, the Russian coreference corpus, used in the Ru-Eval-2014 competition of Russian coreference resolvers. The corpus consists of 180 texts, containing 3638 coreferential chains with the total of 16557 coreferential mentions. The texts of the corpus are of various lengths and genres: fiction, news texts, scientific articles, blog posts, etc. All texts are tokenized and morphologically and syntactically tagged, which allows to use the data without additional preprocessing.

For the evaluation procedure the texts of the corpus were split into training, validation and test datasets in the 60/20/20 proportion.

5.2 Evaluation Results

In our research we compared the performance of two models based on two feature sets, described above: the default one (model I) and the one enhanced with semantic relatedness measures (model II). Several versions of the second model were considered: supplemented with semantic features calculated on only one of the resources, and on both at once.

All models were at first pretrained to determine the proper feature weights, and then their performance on the test dataset was evaluated. For evaluation the MUC [23] and the B^3 [24] metrics were used. The comparison was conducted using the gold mentions from the RuCor corpus. The results of evaluation, as well as results of similar research described in [14] and [16] with the highest B^3 score, are presented in Table 2. The table also includes the absolute error counts for evaluated models.

Table 2. Evaluation results

	MUC			B^3			Error counts		
	P	R	F1	P	R	F1	FN	FA	WL
Model I	0.683	0.607	0.643	0.568	0.644	0.604	217	63	4.8 K
Model II, RuThes only	0.693	0.729	0.710	0.571	0.624	0.597	230	60	4.5 K
Model II, Wikipedia only	0.641	0.679	0.660	0.566	0.659	0.609	0	79	5 K
Model II, both sources	0.693	0.730	**0.711**	0.568	0.682	**0.620**	**100**	**2**	**4.3 K**
[14], random forest	0.740	0.652	0.693	0.739	0.552	***0.631***			
[16], NamedEntities	0.794	0.637	***0.707***	0.794	0.489	0.605			

As can be seen from the table, the general performance of both our models is comparable to that of state-of-the-art systems by [14] and [16]. The MUC score of the variant of Model II that uses the semantic data from both sources is higher than scores of comparison targets, and its B3 score, while 1% lower than that in [14], exceeds the result of [16] by 1.5%. This variant achieved the highest F-measure of all model II variants compared, showing the improvement that can be gained by using semantic features in the analysis. The improvement is also demonstrated by the decrease in error counts of all error types.

Seeing that the difference of our results from the comparison targets is mostly in the lower precision score, increasing it should become the focus of future work.

5.3 Discussion of Results

The results presented above demonstrate that neural networks are viable as a method of coreference resolution in Russian. Trained upon a similar set of features, they perform on par with state-of-the art systems, and only slightly worse than the system using mention clustering. Apart from that, our results show that features derived from semantic information can be successfully used to boost the quality of system's analysis.

Using the features derived from Wikipedia data improves the recall of the system, which can be attributed to large size of the encyclopedia and its coverage of various phenomena. The features derived from thesaurus data, on the other hand, serve to improve the precision, thus the largest increase in quality being gained by combining the features from both information sources. Still, lack of substantial increase in precision after adding semantic features can be observed, which calls for improvements in the feature generation process.

While features based on thesaurus and encyclopedic data help improve coreference resolution for ontologically related mentions, such as hypernyms or synonyms, other complicated cases of coreference still persist. Among them are:

- Direct speech pronouns. First and second person pronouns can be difficult to resolve for a neural network due to their morphological differences from 3rd person ones.
- Split antecedents. Pairs such as *"Иван Тихонович и Татьяна Финогеновна"* ('Ivan Tikhonovich and Tatiana Finogenovna') and *"они"* ('they') will also have differing morphological features, because both heads of the first mention will be analysed separately.
- Relations, depending on context. Cases such as *"Выходец из Нигерии$_i$ решил остаться на ПМЖ в Израиле, поскольку на родине$_i$ его якобы преследует опасный призрак"* ('A native of Nigeria$_i$ decided to remain in Israel, because he was haunted by a ghost in his homeland$_i$') are difficult to resolve, because the understanding of the whole sentence is required to link *"homeland"* to the correct country.

To target such cases, additional improvements need to be made in both the structure of the model and the features used. For example, the context-dependent relations can be possibly resolved by implementing similarity measures between word embeddings.

6 Conclusion

In this paper we presented a neural network, designed for the purpose of coreference resolution in the Russian language, and tested two different feature sets to estimate the importance of semantic features for its performance. The results of evaluation using MUC and B^3 metrics demonstrated that its baseline performance is comparable to that achieved in recent researches on the same topic, and that integration of semantic features helps to increase the quality of analysis to a certain degree.

To target the shortcomings of the system, such as low improvements in precision score, as well as complicated coreference cases, future improvements both in the network architecture and semantic feature extraction process are needed. They include: (i) testing alternative network architectures, including recurrent neural networks; (ii) tuning of hyperparameters; (iii) use of alternative word embeddings and features based on them (the BERT language model is of particular interest); (iv) improving the extraction process of semantic features; (v) testing of other relatedness measures, including between word embeddings. Another important development is integration of clustering and cluster-ranking modules to account for entity-level information.

References

1. Hobbs, J.: Resolving pronoun references. Lingua **44**, 311–338 (1978)
2. Soon, W.M., Lim, D.C.Y., Ng, H.T.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. **27**, 521–544 (2001). https://doi.org/10.1162/089120101753342653
3. Chen, C., Ng, V.: Combining the best of two worlds: a hybrid approach to multilingual coreference resolution. In: Joint Conference on EMNLP and CoNLL-Shared Task, pp. 56–63 (2012)
4. Raghunathan, K., et al.: A multi-pass sieve for coreference resolution. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 492–501. Association for Computational Linguistics (2010)
5. Fernandes, E.R., dos Santos, C.N., Milidiú, R.L.: Latent structure perceptron with feature induction for unrestricted coreference resolution. In: Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task, pp. 41–48 (2012)
6. Clark, K., Manning, C.D.: Improving coreference resolution by learning entity-level distributed representations. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pp. 643–653 (2016). https://doi.org/10.18653/v1/P16-1061
7. Bengtson, E., Roth, D.: Understanding the value of features for coreference resolution. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 294–303. Association for Computational Linguistics, Honolulu, Hawaii (2008)
8. Ponzetto, S.P., Strube, M.: Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 192–199. Association for Computational Linguistics (2006). https://doi.org/10.3115/1220835.1220860
9. Toldova, S., et al.: RU-EVAL-2014: evaluating anaphora and coreference resolution for Russian. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", pp. 681–694 (2014)

10. Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos, S.: A mention-synchronous coreference resolution algorithm based on the bell tree. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004) (2004)
11. Yang, X., Zhou, G., Su, J., Tan, C.L.: Coreference resolution using competition learning approach. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, ACL 2003, pp. 176–183 (2003). https://doi.org/10.3115/1075096.1075119
12. Rahman, A., Ng, V.: Supervised models for coreference resolution. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, vol. 2, pp. 968–977 (2009). https://doi.org/10.3115/1699571.1699639
13. Martschat, S., Strube, M.: Latent Structures for coreference resolution. Trans. Assoc. Comput. Linguist. **3**, 405–418 (2015)
14. Sysoev, A.A., Andrianov, I.A., Khadzhiiskaia, A.Y.: Coreference resolution in Russian: state-of-the-art approaches application and evolvement. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", pp. 341–352 (2017)
15. Versley, Y.: Antecedent selection techniques for high-recall coreference resolution. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007)
16. Toldova, S., Ionov, M.: Coreference Resolution for Russian: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", pp. 339–348 (2017)
17. Azerkovich, I.: Employing Wikipedia data for coreference resolution in Russian. In: Filchenkov, A., Pivovarova, L., Žižka, J. (eds.) AINL 2017. CCIS, vol. 789, pp. 107–112. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-71746-3_9
18. Loukachevitch, N.V., Dobrov, B., Chetviorkin, I.: RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014)
19. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. **19**, 17–30 (1989). https://doi.org/10.1109/21.24528
20. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: ACL, pp. 133–138 (1994). https://doi.org/10.3115/981732.981751
21. Leacock, C., Chodorow, M.: Combining local context with wordnet similarity for word sense identification. In: WordNet: An Electronic Lexical Database, pp. 265–283. MIT Press (1998)
22. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. IJCAI 448–453 (1995)
23. Vilain, M., Burger, J.D., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6), pp. 45–52 (1995). https://doi.org/10.3115/1072399.1072405
24. Baldwin, B., Bagga, A.: Algorithms for scoring coreference chains. In: The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, pp. 563–566 (1998)