Using Thesaurus Data to Improve Coreference Resolution for Russian

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

Abstract

Semantic information about entities, specifically, how close in meaning two mentions are to each other, can become very useful for the task of coreference resolution. One of the most well-researched and widely used forms of presenting this information are measures of semantic similarity and semantic relatedness. These metrics are often computed, relying upon the structure of a thesaurus, but it is also possible to use alternative resources. One such source is Wikipedia, which possesses the category structure similar to that of a thesaurus. In this work we describe an attempt to use semantic relatedness measures, calculated on thesaurus and Wikipedia data, to improve the quality of a coreference resolution system for Russian language. The results show that this is a viable solution and that combining the two sources yields the most gain in quality.

1 Introduction

Coreference resolution is a very important part of many natural language processing tasks, and for solving it generally information from several language layers is required. Among those, the importance of semantic information, as opposed to more shallow features, e.g. string-based, morphologic or syntactic ones, is sometimes debated (see e.g. Durrett and Klein (2013)), but it is nevertheless seen as useful for overcoming the potential plateau of quality, as V. Ng (2017) noted.

As far as English language is concerned, various thesauruses are usually used as sources of semantic information, the most popular of them being the WordNet (Harabagiu et al., 2001; Ponzetto and Strube, 2006 among others). Another such resource is Wikipedia that, while not a thesaurus by itself, is sometimes considered as such due to its structure of categories, connected to each other by the relation of inclusion (Ponzetto and Strube, 2006).

For Russian language the room for improvement of coreference resolution systems still exists, as has been demonstrated by results of the Ru-Eval-2014 competition for Russian coreference resolvers (Toldova et al., 2014). The usage of semantic information is also not as widespread, partly due to lesser volume of resources available: fewer thesauruses exist for Russian than there are for English, the most prominent of them being the RuThes (Loukachevitch et al., 2014), consisting of appr. 70 000 synsets, and the Russian segment of Wikipedia is also smaller. Consequently, fewer attempts at using semantic information have been made.

Nevertheless, the results of Toldova et al. (2014) mentioned above clearly show that semantic information needs to be explored to properly resolve cases such as (1) below.

(1) People who survived the wreck of **the ship** told that the main reason for the tragedy was the **oil-burner** being very old.

Additional information that can be obtained from a thesaurus is required to correctly join *oil-burner* to *the ship*. On the other hand, while thesauruses seldom contain information about named entities, such as people, additional resources would be required to obtain information of this kind. Data that can only be obtained from an encyclopedia such as Wikipedia is required for examples like (2):

(2) Victor Vekselberg would like to engage **Grigori Perelman** to work in the "Silicon Valley". The fortune has smiled upon **the mathematician**…

To deal with cases similar to the ones described above, a system would require to look-up the related content in a resource and properly infer the relation between the mentions.

This paper presents an attempt at using information, obtained from RuThes and Russian Wikipedia, to improve the quality of coreference resolution for the Russian language. More precisely, we explore the efficiency of using measures of semantic similarity and semantic relatedness, as quantified representations of how close the meanings of two concepts are. In our research we employ the measures, extracted from the aforementioned resources, as features used in machine learning solutions.

The achieved results suggest that integrating features based on semantic information does indeed improve the system performance, with the highest increase in quality being gained by combining the data from both resources.

2 Related Work

Thesauruses, in particular WordNet, have been widely used for purposes of coreference resolution in a variety of ways. Some of these include extracting hypernym chains or semantic classes, derived from high-level nodes (Poesio and Vieira, 2000; Soon et al., 2001) or calculating special confidence measures of different paths between concepts (Harabagiu et al., 2001). Semantic similarity has also been frequently employed in automated coreference resolution, either calculated from thesaurus data or unannotated corpora (Ponzetto and Strube, 2006; Versley, 2007), or based on word embeddings (Clark and Manning, 2016). A large spectrum of different semantic similarity values that can be calculated based on thesaurus structure has been suggested by various researchers. Overview of the most influential ones are given, e.g., in (Budanitsky and Hirst, 2006).

For Russian the research of coreference resolution using thesaurus data has been smaller in scale with the only participant system of RuEval-2014 that used semantic information relying on a proprietary ontology (Bogdanov et al., 2014). Recently, Toldova and Ionov (2017) have introduced a coreference resolution system, supplemented with semantic information from hypernym chains extracted from RuThes, achieving certain improvements in quality. Our research differs in approach with employing semantic similarity measures instead.

The Wikipedia data is also often used in systems of coreference resolution, including the Stanford parser (Raghunathan et al., 2010). Generally, the text content of the page is considered for analysis, with its category structure being used in a similar way to a thesaurus in (Ponzetto and Strube, 2006). The text information and categories of a page from Russian Wikipedia have been used by Azerkovich (2018) with a positive result, but the category tree as a whole was not considered.

3 Calculating Semantic Relatedness

3.1 Resources Used

Two main sources of semantic information were used in this research: RuThes thesaurus and the Russian segment of Wikipedia. RuThes is a thesaurus, created by a team of linguists, with its freely available part, RuThes-Lite, including 55 000 entities that correspond to 158 000 lexical entries. The structure of RuThes is similar to that of WordNet, with concepts in the thesaurus linked to each other by the set of labeled relations that includes IS-A, PART-WHOLE and a number of associative relations.

The Russian segment of Wikipedia with ~1.5 mln articles, while being smaller than the English one (over 5 mln articles), is still one of its largest, making it an important knowledge source. The feature of Wikipedia that allowed to include its information in our analysis is its category structure: each article can be placed within one or several categories, which, in its own turn, can be categorized further. Because one article can belong to several categories, and one category can be included in several parent categories, the structure of Wikipedia categories is not a tree in a strict sense, but a more general graph.

For both resources the following set of measures of semantic similarity was calculated: the path-based measures of Rada et al. (1989), Wu and Palmer (1994) and Leacock and Chodorow (1998); information content-based measure of Resnik (1995). Because the relations between parent and child categories in Wikipedia do not strictly correspond to IS-A relations, it would be more correct to consider the scores for this source as measures of semantic relatedness rather than semantic similarity.

For Wikipedia pages the measure of gloss overlap by Banerjee and Pedersen (2003) was also computed. This was not done for RuThes data, because not all synsets there are provided with a gloss, which is required to apply this measure.

3.2 Mining Semantic Information

In the case of RuThes, values of semantic similarity measures for two referential expressions

were obtained by calculating the scores for head lemmas of the groups in question. In case of heads of any or both groups being ambiguous, measures for all possible combinations of meanings were obtained, and after obtaining the values, the following two features were created: the maximum value of the similarity score, and the average value of the similarity score. If one or both mentions were absent from the thesaurus, the measure scores were considered to be zero.

In the case of Wikipedia, the problem of ambiguity had to be addressed slightly differently. To calculate the semantic relatedness measures, firstly, the pages corresponding to the referring expressions in question had to be obtained. For that purpose, the groups were queried to Wikipedia search engine. In case a disambiguation page was encountered, all hyperlinks from the page were analyzed. If a link led to the page, containing the other queried group, it was used as the hit. If no such links were found, the first hyperlink on the page was used. After resolving the referring expressions to their Wikipedia pages, the gloss overlap measure of the pages' texts was calculated.

The rest of the set of metrics was calculated in the same way as for RuThes, using the graph of categories to which the obtained pages belong. Following the observations of Ponzetto and Strube (2006), the possible depth of nodes was limited to 4 to assure less noisy results, due to higher levels of the category structure being too strongly connected. The values of path-based and information content-based measures were obtained for all combinations of categories for both pages, after which the same two features as for thesaurus data was calculated: the maximum value of the similarity score, and the average value of the similarity score. As with the RuThes data, if any of the mentions was not mapped to a corresponding Wikipedia page, the measures were considered zero.

3.3 Correlation with Human Judgement

As an additional step in preparing to use the values of measures, described above, as features for a coreference resolution algorithm, it was tested to what extent these measures correlate with human judgement on coreference.

To achieve that, the chosen set of measures was calculated for a set of referring expressions with pre-existing coreference annotation. As the source of annotation, the Russian coreference corpus RuCor was used. It is the corpus, created for the purposes of the task of automated anaphora and coreference resolution for RU-EVAL-2014 (Toldova et al., 2014). For 200-pair sets of coreferent and non-coreferent pairs semantic relatedness was calculated, and then the Pearson correlation coefficient with the annotation was calculated. To enable the calculations, the pairs from the evaluation set were assigned the maximum measure value if they were annotated as coreferent, and the minimum value if marked as not coreferent.

The results of evaluation are presented in Table 1. As can be seen from the tables, the values of measures generally do correlate with human judgement, justifying their usage as features for analysis, except from the gloss overlay, which was not used in further experiments. Different measures also correlate differently with coreference annotation: while the measures, obtained on the data from RuThes display higher correlation in general, the data from Wikipedia correlates relatively well with annotation for named entities. This leads to conclude that combining data from both resources can give the most coverage and, potentially, a larger improvement to quality of the analysis.

Source	***Rada***	***Wu***	***Leacock***	***Resnik***	***Gloss***
RuThes (non-empty)	0.56	0.59	0.51	0.30	n/a
Wikipedia (NEs)	0.7	0.6	0.1	0.2	0.2

Table 1: Correlation with coreference annotation

4 Using Semantic Relatedness for Machine Learning Feature Creation

4.1 Corpus Data Used

The research was conducted on the data of the aforementioned RuCor corpus (Toldova et al., 2014), as the largest available corpus of Russian with coreference annotation. It consists of 180 texts of a variety of genres that in total contain 3838 coreferential chains with 16557 referential expressions. For the Ru-Eval-2014 task it was split in the training and test sets (70% and 30% of the corpus volume, respectively), which were retained for our experiments. All texts in the corpus have been preprocessed and morphologically tagged using the set of instruments developed by Sharoff and Nivre (2011). The annotation, provided by the corpus creators, was used as the

golden standard, against which the systems were evaluated.

4.2 Learning Algorithm

For our research we used a machine learning algorithm based on a decision tree classifier, which has been tested in application to coreference resolution for Russian in (Toldova and Ionov, 2017). It is based on the work of (Soon et al., 2001), and uses a similar set of baseline features that we supplemented with described above features, derived from thesaurus data.

The system is based on a pairwise approach, according to which the classifier, being given a pair of referring expressions, decides whether they corefer or not, based on the feature values. The candidate pairs for analysis were created the following way: from each pair of coreferent expressions a positive instance is created, and then every NP between the anaphor and the antecedent is paired to the anaphor to create a negative instance. In our research we relied upon the NP boundaries, obtained from the corpus markup instead of automatically generated ones, in order to maximize the influence of the features we introduce in addition to the baseline set.

4.3 Baseline Features

The baseline system was based on the set of features, derived from the original set, suggested by Soon et al. (2001). It included features of various types: string-based, distance, morphological, syntactic and semantic. But, as it was originally created for the English language, several features, such as definiteness, were meaningless in the case of Russian, due to linguistic differences. Because of that, they were removed and, in some cases, replaced with alternative ones. The resulting feature set is given in Table 2.

Feature type	Features
String features	• Mention strings match • One of mentions is an identifier of the other • One of mentions is an abbreviation of the other
Distance features	• Number of sentences between mentions • Number of sentences is greater than 3
Morphological features	• Mentions match in gender • Mentions match in number • Both mentions are proper • Anaphor is a demonstrative pronoun • One of mentions is a pronoun
Syntactic features	• The potential anaphor is an appositive of the antecedent • Mentions are subject and object of the same sentence • Both mentions are subjects • Both mentions are first words in a sentence
Semantic features	• Both mentions are animate

Table 2: Baseline feature set

All features were represented by their numeric value if applicable, or indicator functions, equal to 1 in case the feature is true, and 0 in case it is false.

The performance of the system, using only the baseline set, was compared to performance of its version, using the set enhanced with features derived from thesaurus data of RuThes and Wikipedia: maximum and average values of the semantic relatedness measures.

4.4 Performance Evaluation

The performance of systems was evaluated, based upon a number of metrics: MUC (Vilain et al., 1995), B^3 (Baldwin and Bagga, 1998) and CEAF (Luo, 2005). The following versions of the baseline system were included in the comparison: enhanced with the RuThes-based features; enhanced with Wikipedia-based features; enhanced with features from both resources.

The Table 3 below contains the results of the comparison by metric, with maximum improvements over the baseline highlighted in bold. The improvements, achieved in the aforementioned work of Ponzetto and Strube (2006) by adding Wikipedia-based and Wordnet-based features are also given for comparison.

4.5 Discussion

The results of the evaluation show that features based on semantic relatedness measures do increase the system performance compared to the baseline to a certain degree. While the increase is similar in scale to the numbers demonstrated in earlier work of Ponzetto and Strube (2006), it may still be not large enough for statistical importance This prevents us from labelling it a decisive improvement and calls for further development of the method.

	MUC			B^3			CEAF
	P	R	F	P	R	F	
Baseline	72.76	59.49	65.46	71.01	44.50	54.71	49.02
Baseline + Wikipedia	70.28	59.71	64.56	66.50	44.63	53.41	46.36
Baseline + RuThes	72.72	59.43	65.41	71.15	44.44	54.71	48.91
Baseline + RuThes + Wikipedia	73.57	60.01	**66.10**	71.77	44.93	**55.26**	**49.66**
(Ponzetto and Strube, 2006), Wikipedia	+1.3%	-0.5%	+0.8%				
(Ponzetto and Strube, 2006), Wordnet	+2.2%	-0.9%	+1.3%				

Table 3: Evaluation metrics

Still, the resulting increase in quality is larger compared to that of similar work by Toldova and Ionov (2017): 0.54% of MUC score and 0.55% of B^3 score, compared to 0.26% and 0.19% correspondingly. As in our research we used semantic information in the form of semantic relatedness measures, compared to hypernym chains in (Toldova and Ionov, 2017), we can assume that more precise preprocessing of information and usage of features beyond Boolean ones can lead to more improvements in systems' performance.

Study of the results reveals that the largest increase in quality is observed when combining the features from both sources, with the improvement seen across all evaluation metrics. This corresponds to the assessment of correlation with human judgement described above.

The results also allow to conclude that information from both used sources serves to improve the quality of the analysis in different ways. While the data from RuThes can be used to improve the system's precision, the data from Wikipedia helps to increase the recall of the performance. This can be contributed to the difference in content between the sources: while RuThes, as a thesaurus created by a team of linguists, is less in size, but better structured than Wikipedia, the latter possesses a more contrived and not necessarily transparent category system, but contains more information about wider range of phenomena.

5 Conclusions and Future Work

In this paper we described an attempt to improve the quality of coreference resolution for Russian by introducing features, based on semantic information, obtained from thesaurus data. For that end, we used the thesaurus of Russian RuThes and the Russian segment of Wikipedia to compute several semantic relatedness measures to be used as features in a coreference resolution system.

While the results of evaluation of the system cannot yet be called final, they suggest that the quality of coreference resolution for Russian can be improved by using features based on semantic information. It is important to remark that the maximum profit was achieved by combining the features from both sources, with Wikipedia also being useful despite its open-source nature and being open to free editions by any user. While recent research relying on neural networks for coreference resolution achieve better results for Russian (e.g. (Le et al., 2019)), the gains of using semantic information observed by us and other researchers allow to assume that such algorithms could benefit from implementing it, as well.

Future work, inspired by this research, lies in exploring other coreference resolution algorithms and improving the quality of semantic features extraction. The former involves exploring more productive techniques of coreference resolution, in particular, assessing the potential of integrating semantic level information in neural networks. The latter involves employing a wider range of semantic relatedness measures, as well as increasing the efficiency of using Wikipedia-based information. As an alternative to the online encyclopedia, DBpedia can be used. It possesses clearer structure and labeled relations, which could simplify computing semantic relatedness from its data.

References

Ilya Azerkovich. 2018. Employing wikipedia data for coreference resolution in Russian. In *Artificial Intelligence and Natural Language*, volume 789, pages 107–112. Springer, Cham.

Breck Baldwin and Amit Bagga. 1998. Algorithms for Scoring Coreference Chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805--810.

A. V. Bogdanov, S. S. Dzhumaev, D. A. Skorinkin, and A. S. Starostin. 2014. Anaphora analysis based on ABBYY Compreno linguistic technologies. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014)*, volume 13, pages 89–102.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, March.

Kevin Clark and Christopher D. Manning. 2016. 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)*, pages 643–653. June.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982.

Sanda M. Harabagiu, Razvan C. Bunescu, and Steven J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *2nd Meeting of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 55–62. The Association for Computational Linguistics.

T. A. Le, M. A. Petrov, Y. M. Kurato, and M. S. Burtsev. 2019. Sentence Level Representation and Language Models in The Task of Coreference Resolution for Russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2019)*, pages 341–350.

Claudia Leacock and Martin Chodorow. 1998. Combining local context with wordnet similarity for word sense identification. In *WordNet: an electronic lexical database*, pages 265–283. MIT Press.

Natalia V. Loukachevitch, Boris Dobrov, and Ilia Chetviorkin. 2014. 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)*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.

Vincent Ng. 2017. Machine Learning for Entity Coreference Resolution : A Retrospective Look at Two Decades of Research. In *Proceedings of the 31th Conference on Artificial Intelligence (AAAI 2017)*, volume 6, pages 4877–4884.

Massimo Poesio and Renata Vieira. 2000. An Empirically Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4):539–593, December.

Simone Paolo Ponzetto and Michael Strube. 2006. 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*, volume 33, pages 192–199. Association for Computational Linguistics.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *IJCAI*:448–453.

S. Sharoff and J. Nivre. 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2011)*, volume 10, pages 591–605.

Wee Meng Soon, Daniel Chung Yong Lim, and Hwee Tou Ng. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, December.

Svetlana Toldova and M. Ionov. 2017. Coreference Resolution for Russian: The Impact of Semantic Features. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2017)*, volume 16, pages 339–348.

Svetlana Toldova, Anna Roitberg, Alina Ladygina, M. D. Vasilyeva, Ilya Azerkovich, M. Kurzukov, Galina Sim, D. V. Gorshkov, A. Ivanova, Anna Nedoluzhko, and Yulia Grishina. 2014. RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014)*, volume 13, pages 681–694.

Yannick Versley. 2007. 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*.

Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52.

Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. *ACL*:133–138.