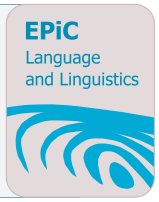
EPiC Series in Language and Linguistics

Volume 4, 2019, Pages 26–32

Proceedings of Third Workshop "Computational linguistics and language science"



Evaluating measures of semantic relatedness for Russian language

Ilya Azerkovich[1]

[1] National Research University Higher School of Economics, Moscow, Russia

ilazerkovich@edu.hse.ru

Abstract

Coreference resolution is recognized as an important task in natural text processing and it has been proven that knowledge of semantic relations between two possibly coreferent entities can provide a certain increase in quality for automated solutions. One of the ways to integrate semantic information in such a system is to measure semantic relatedness between candidates for establishing coreference relation. This research is devoted to evaluating the efficiency of different types of semantic relatedness metrics, calculated from different sources, for coreference resolution on the material of Russian language.

1 Introduction

Coreference resolution is a very important part of many natural language processing (NLP) tasks, which generally requires information from several language layers. As a rule, morphological and syntactical information is used, but as of late researchers have been pointing out the importance of integrating semantical information in the process of solving this task ((Azerkovich, 2018), (Ponzetto & Strube, 2006), (Rahman & Ng, 2011), (Toldova & Ionov, 2017)). One of the most transparent ways of representing semantic information for potential use in automated coreference resolution systems are measures of semantic relatedness between entities. These measures have also been successfully used in recommendation systems, bioinformatics do-main and for other NLP tasks. While originally calculated on and applied to taxonomy-my data, semantic relatedness has also started to be applied to such a promising source of information as Wikipedia ((Gabrilovich & Markovitch, 2007), (Seco, Veale, & Hayes, 2004)).

Research devoted to enhancing algorithms of coreference resolution with semantic information is also being conducted on the material of Russian language ((Azerkovich, 2018), (Toldova & Ionov, 2017)). The resources of semantic information for Russian language are less in number and extensiveness than for English (~1.5mln vs ~5.6mln Wikipedia articles in corresponding language segment, or ~70 000 synsets vs ~117 000 in corresponding Wordnets –RuThes (Loukachevitch, 2011) for Russian, Princeton Wordnet for English), but nevertheless the re-search has proven that quality of coreference resolution can be considerably improved in this way. As the main goal of my ongoing

G. Wohlgenannt, R. von Waldenfels, S. Toldova, E. Rakhilina, D. Paperno, O. Lyashevskaya, N. Loukachevitch, S.O. Kuznetsov, O. Kultepina, D. Ilvovsky, B. Galitsky, E. Artemova and E. Bolshakova (eds.), CLLS 2018 (EPiC Series in Language and Linguistics, vol. 4), pp. 26–32

research is integrating semantic information in systems of coreference resolution for Russian, an attempt to implement semantic relatedness features was a natural step.

This work presents one stage of the research, dedicated to evaluating existing metrics of semantic relatedness for coreference resolution in Russian. It describes the results that were achieved by computing a set of metrics using two different sources, RuThes and Wikipedia, as freely available and considerably large collections of se-mantic information for the Russian language. The comparison of the metrics against each other, as well as against human judgement demonstrated that while metrics based on RuThes data proved to be more reliable in general, measures from Wikipedia data demonstrated considerable precision for pairs containing named entities. Future work based on this research includes employing these metrics as features in a machine learning-based system of automated coreference resolution.

2 Related Work

A lot of research has been done in the field of semantic relatedness, and a considerable number of different metrics has been suggested to this day. In general, computing semantic relatedness between entities using data from a lexicographic resource is done by representing this resource as a graph and examining paths within it.

As a rule, semantic relatedness measures are obtained on data from ontologies ((Resnik, 1995), (Wu & Palmer, 1994)), but with the appearance and growing popularity of Wikipedia as a source of information, semantic relatedness measures have also started adapting to its structure. In (Strube & Ponzetto, 2006) adjustments are suggested for several such metrics to better account for category structure and length of Wikipedia articles. Wikipedia derivatives, such as Dbpedia, are also frequently used (e.g. (Leal, Rodrigues, & Queirós, 2012) describes a recommendation system built upon information from this source).

Semantic relatedness calculated on web data has been used for such NLP tasks as entity disambiguation (Bunescu & Marius, 2006) or coreference resolution (Rahman & Ng, 2011). Attempts at using semantic information for coreference resolution in Russian language have also been done, in the form of gazetteers in (Toldova & Ionov, 2017) or Wikipedia articles in (Azerkovich, 2018). This work hopes to demonstrate that quality of semantic information, described in these works, can be further improved upon by using semantic relatedness measures calculated on free open data sources.

3 Semantic Relatedness Measures

Semantic measures, considered in this research, can be grouped in three large classes: 1) path-based measures; 2) measures that are calculated as a function of information content between two entities; 3) measures, based on gloss overlaps between definitions.

3.1 Path-based Measures

These measures are calculated from number of edges of the path in the ontology representation between the two concept nodes c_1 and c_2, corresponding to the words w_1 and w_2 in question. Semantic relatedness is then defined as the inversion of the path measure. The simplest such metric would be the count of the edges along the shortest path between the two nodes (pb_1), suggested in (Rada, Mili, Bicknell, & Blettner, 1989). A normalization method that takes into account the total depth of the ontology containing the concepts has been suggested in (Leacock & Chodorow, 1998) (pb_2)

$$pb_2(c_1, c_2) = -\log\frac{length(c_1,c_2)}{2D}, \quad (1)$$

where *length(c_1, c_2)* corresponds to metric pb_1 mentioned above, and D is the maximum depth of the ontology.

Another modification (pb_3), suggested by (Wu & Palmer, 1994), involves scaling by the depth of the least common superconcept (*lcs*) node of c_1 and c_2, apart from depths of c_1 and c_2.

$$pb_3(c_1, c_2) = \frac{2*depth(lcs)}{length(c_1,lcs)+length(c_1,lcs)+2*depth(lcs)} = \frac{2*depth(lcs)}{depth(c_1)+depth(c_2)}, \quad (2)$$

where *depth(node)* represents distance from root of the taxonomy to the node.

3.2 Information Content-based Measures

In (Resnik, 1995) semantic relatedness is interpreted as a measure of information content (*ic*) of the least common superconcept of the two concepts in question. Probability of the concept is calculated from occurrences of corresponding words in a corpus.

Strube and Ponzetto in (Strube & Ponzetto, 2006) suggest reinforcing this metric with intrinsic information content measure from (Seco, Veale, & Hayes, 2004) instead of word frequencies, as it better correlates with human judgement. Intrinsic information content is calculated from the number of hyponyms of the node

$$ic = 1 - \frac{\log(hypo(lcs)+1)}{\log(max_{wn})}, \quad (3)$$

where max_{wn} is the total number of nodes in the ontology.

3.3 Text Overlap-based Measures

These metrics are defined as a function over text (gloss, or definition) overlaps (*to*), suggested in (Lesk, 1986) for dictionary definitions. A variation of such measure is the *extended gloss overlap*, suggested in (Banerjee & Pedersen, 2003). It is suggested to calculate the overlaps with the use of the following formula: $overlap(t_1, t_2) = \sum_n m^2$, for n m-word overlaps between texts t_1 and t_2. In the case of Wikipedia, the first paragraph of the article is considered to be the gloss, as well as the full text of the article. Normalization targeted at minimizing effects of different text lengths, introduced in (Strube & Ponzetto, 2006), is also used.

$$to(t_1, t_2) = \tanh\left(\frac{overlap(t_1,t_2)}{length(t_1)+length(t_2)}\right) \quad (4)$$

In this equation *length(text)* represents length of the text in question, not the path within the ontology.

4 Calculating Semantic Relatedness

For the purposes of this research two sources of semantic information were considered: Russian segment of Wikipedia, and a Russian Wordnet, RuThes-lite 2.0 ((Loukachevitch, 2011)). They were chosen as freely available sources of information that were the most complete at the time this work was being written.

The hierarchic structure of RuThes is transparent and explicitly described. It contains concepts, connected by a set of relations: hyponym/hypernym relation, meronym/holonym relation and associative relations. For the purposes of this paper, only the hyponym/hypernym relation was considered, as it better reflects the coreference relation, addressed further in the paper.

On the other hand, additional search among Wikipedia categories was required to calculate path-based and informational content-based measures. As RuThes entries contained no definitions, and

only taxonomic information, the *to* metric was calculated only for Wikipedia data. Named entities are also absent from the thesaurus, so if they were present in a word pair in question, only Wikipedia-based metrics could be calculated for such a pair, as well.

4.1 Obtaining and disambiguating Wikipedia pages

Wikipedia pages $p_{i/j}$, corresponding to concepts $c_{i/j}$ were obtained by querying corresponding words *i/j* to Wikipedia search engine. In case a disambiguation page $p_{i/j}$ was returned, the following course of action was taken. Namely, the other member of the pair, *j/i*, as well as all hyperlinks from its corresponding page *i/j* were compiled in a list of possible disambiguating terms. E.g., for the pair of terms *<Гугл 'Google', поисковик 'search engine'>*, querying the first member yields a disambiguation page, containing links to Google as a search engine, Google as a company, etc. Then a list containing items {*поисковая система 'search engine', информационный поиск 'information retrieval', веб-служба 'web service',* etc.} is created. Next, if a link on page $p_{i/j}$ contains an item from the list, the linked page is returned, otherwise the first link on $p_{i/j}$ is returned. In the case discussed above, the term 'search engine' is contained in one of the links from the disambiguation page, and so the page titled *"Google (поисковая система)" 'Google (search engine)'* is correctly chosen.

Entity disambiguation is a complex task by itself, and a lot of research is dedicated solely to solving it, but the method described here produced plausible results while being relatively fast and non-consuming.

4.2 Obtaining paths from category structure

After disambiguating Wikipedia pages, from each of them a complete set of categories was obtained. The set of categories was considered the same as nodes of the thesaurus, and belonging to a category the same as an "is-a" relation. Accordingly, links between categories were followed until the least common superconcept was found. This allowed to calculate the same path-based and information content-measures for Wikipedia, as for RuThes.

5 Evaluation

Entities for calculating semantic relatedness measures were obtained from Russian coreference corpus RuCor, created for the task of automated anaphora and coreference resolution for Russian RU-EVAL-2014 (Toldova, et al., 2014). It was chosen because it already contains coreference markup, i.e. provides pairs of entities with annotated semantic relatedness beforehand, without the need for additional human annotation.

In total the corpus provides almost 800 sets of mentions, or coreferential chains, containing two or more noun phrases or named entities. From these chains the set of 200 pairs of coreferent entities and 200 pairs of not coreferent entities was formed. Only non-coinciding noun phrases were included, to ensure that calculated similarity metrics would be meaningful for all pairs in the set.

Then for each pair of entities from the evaluation set the following semantic relatedness metrics were calculated, based on RuThes and Wikipedia data: *pb1, pb2, pb3, ic* and *ego*. All groups in the corpus have their heads marked, so for multiword ex-pressions the marked heads were considered for calculating RuThes metrics, while whole expressions were used for search in Wikipedia.

As a baseline, Jaccard similarity coefficient was calculated for each word pair *i* and *j*, based on number of Google hits:

$$jaccard = \frac{hits(i\ and\ j)}{hits(i)+hits(j)-hits(i\ and\ j)} \quad (5)$$

Gold standard results were obtained as follows: depending on the metric, the pairs from the evaluation set were assigned the maximum metric value if they were marked as coreferent, and the minimum value if marked as not coreferent. Then for each of the calculated metrics the Pearson correlation coefficient with the gold standard for the metric was calculated. The results are presented in Tables 1 and 2 below.

	Baseline	RuThes				
	jaccard	*pb1*	*pb2*	*pb3*	*ic*	*to*
all	0.34	0.28	0.38	0.34	0.16	n/a
non-missing	0.34	0.56	0.59	0.51	0.30	n/a

Table 1: Correlation with human judgment for calculated measures for RuThes

	Baseline	Wikipedia				
	jaccard	*pb1*	*pb2*	*pb3*	*ic*	*to*
all	0.34	0.05	0.35	0.58	0.23	0.03
named entities	0.6	0.7	0.6	0.08	0.2	0.2

Table 2: Correlation with human judgment for calculated measures for Wikipedia

As can be seen from the tables, different metrics prove to correlate more with hu-man judgement, depending on the data source. While, as has been noted before, named entities are absent from RuThes, and metrics for them could not be calculated, metrics for entities present in the thesaurus were on average over the baseline and more representative than metrics based on Wikipedia. This can probably be explained by Wikipedia category structure being more fine-grained, which leads to short paths also existing between less related entities. On the other hand, Wikipedia metrics values for named entities are generally higher than for all entities and two of three path-based metrics are on par or higher than the baseline.

Path-based measures demonstrated highest correlation with human judgement among all measures, except *pb3* for Wikipedia-based named entities, which might be due to an unaccounted-for skew in the source data. Measures, based on information content were not informative both for RuThes and Wikipedia, compared to path-based metrics, which partly coincides with the results, obtained in (Strube & Ponzetto, 2006). Specifics of Wikipedia structure that decrease the efficiency of this metric might be a point of another study. Measures, based on text overlaps, were the least effective for Wikipedia for glosses as well as for full texts of the articles. One of the reasons for such low performance might be the style of writing in Russian Wikipedia not supporting repetitions of large enough text parts in different articles. Another aspect of the problem might be excessive normalization of the results, taking into consideration large sizes of some articles paired with low overlap count.

6 Conclusions

In this work we compared a set of metrics of semantic relatedness between pairs of entities, based on paths between nodes, corresponding to the entities in an ontology, informational content of the nodes, and text overlaps between their glosses and full texts of definitions. The metrics were calculated from data obtained from RuThes and Wikipedia as largest and most competent free data sources for Russian.

It was observed that while performance of RuThes-based metrics was in general higher than those based on Wikipedia, the latter performed on or above the baseline in that case. This can be interpreted as the structure of a thesaurus better corresponding to human understanding than structure of Wikipedia categories. While contrasting to results of other related research on larger datasets, e.g. (Strube & Ponzetto, 2006), this may hint at categories of Russian Wikipedia segment being more

confusing than helpful. Wikipedia-based measures may partly perform better for named entities, because pairs including them that are judged as semantically related usually contain their characteristic, occupation (for people) or other feature that is likely to belong to a category closely connected to that entity.

Main direction of future research based on the results of this work is employing the metrics described here for coreference resolution for Russian language. As it has been proven that information from both RuThes and Wikipedia does reflect human judgment in evaluating semantic relatedness for coreferent entities, it makes sense to implement metrics described above as machine learning features for coreference resolution algorithms. Another area of research concerns Wikipedia-based metrics and improvements that need to be done to make these metrics more representative in general, and not only for named entities. Most likely, adjustments need to be done to the process of traversing the category tree to choose the optimal path between nodes. Metric based on text overlaps discussed here should also be revised, because as of now it suffers from too low representativeness.

References

Azerkovich, I. (2018). Employing Wikipedia data for coreference resolution in Russian. (A. Filchenkov, L. Pivovarova, & J. Žižka, Eds.) *Artificial Intelligence and Natural Language. AINL 2017. Communications in Computer and Information Science, 789*.

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *IJCAI (Vol. 03)*, (pp. 805–810).

Bunescu, R., & Marius, P. (2006). Using encyclopedic knowledge for named entity disambiguation. *Proceedings of the 11th conference of the European chapter of the Association for Computational Linguistics (EACL-06)*, (pp. 9-16). Trento, Italy.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJcAI (Vol. 7)*, (pp. 1606-1611).

Leacock, C., & M., C. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet. An Electronic Lexical Database* (pp. 265–283). MIT Press.

Leal, J., Rodrigues, V., & Queirós, R. (2012). Computing semantic relatedness using Dbpedia. *OASIcs-OpenAccess Series in Informatics (Vol. 21)*.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 5th Annual Conference on Sys-tems Documentation* (pp. 24-26). ACM.

Loukachevitch, N. (2011). *Thesauri in Information Retrieval Tasks [publication in Russian].* Moscow: Moscow State University Publishing House.

Ponzetto, S. P., & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. *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.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric to semantic nets. *IEEE Transactions on Systems, Man and Cybernetics, 19*(1), 17-30.

Rahman, A., & Ng, V. (2011). Coreference resolution with world knowledge. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1* (pp. 814-821). Association for Computational Linguistics.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proc. of IJCAI-95, Vol. 1*, (pp. 448–453).

S., T., & M., I. (2017). Coreference resolution for Russian: the impact of semantic features. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"* (pp. 339-348). Moscow: Publishing Center RSUH.

Seco, N., T., V., & J., H. (2004). An intrinsic information content metric for semantic similarity in WordNet. *Proc. Of ECAI-04*, (pp. 1089-1090).

Strube, M., & Ponzetto, S. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. *AAAI, vol. 6*, (pp. 1419-1424).

Toldova, S. J., Roytberg, A., Ladygina, A. A., Vasilyeva, M.D., Azerkovich, I. L., . . . Grishina, Y. (2014). Ru-Eval-2014: Evaluating anaphora and coreference resolution for Russian. *Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2014"* (pp. 681-694). Moscow: Publishing Center RSUH.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *Proc. of ACL-94*, (pp. 133–138).