Use of Neighbor Sentence Co-occurrence to Improve Word Semantic Similarity Detection

Natalia Loukachevitch,
Lomonosov Moscow State University,
Bauman Moscow State Technical University
Moscow, Russia
louk_nat@mail.ru

Aleksey Alekseev
Lomonosov Moscow State University
Moscow, Russia
a.a.alekseevv@gmail.com

Abstract—**In this paper we present the first results of detecting word semantic similarity on several Russian semantic word similarity datasets including the Russian translations of Miller-Charles and Rubenstein-Goodenough sets, the similarity and relatedness subsets of the Russian translation of WordSim353 set prepared for the first Russian word semantic evaluation Russe-2015. The experiments were carried out on three text collections: Russian Wikipedia, a news collection, and their united collection. We found that the best results in detection of lexical paradigmatic relations were achieved using the combination of word2vec with the new type of features based on word co-occurrences in neighbor sentences.**

I. Introduction

Currently, natural language technologies are actively used for processing large volumes of news texts and social network messages. Such tasks as word sense disambiguation, text clustering, textual entailment, search query expansion, etc. often require knowledge about semantic similarity between words.

To check the possibility of current methods to automatically determine word semantic similarity, various datasets have been created. Most known gold standards for this task include the Rubenstein-Goodenough (RG) dataset [1], the Miller-Charles (MC) dataset [2] and WordSim353 [3]. These datasets contain word pairs annotated with human subjects according to lexical semantic similarity. Results of automatic approaches are compared with the datasets using Pearson or Spearman's correlations [4].

Some researchers distinguish proper similarity between word senses that is mainly based on paradigmatic relations (synonyms, hyponyms, or hyperonyms), and relatedness that is established on other types of lexical relations. Agirre et al., [4] subdivided the WordSim353 dataset into two subsets: the WordSim353 similarity set and the WordSim353 relatedness set. The former set consists of word pairs classified as synonyms, antonyms, identical, or hyponym-hyperonym and unrelated pairs. The latter set contains word pairs connected by other relations and unrelated pairs.

To provide word similarity research in other languages, the English similarity tests have been translated [7], [8], [9]. Gurevych translated the MC and RG datasets into German [7], Hassan and Michalchea translated them into Spanish, Romanian and Arabic [8]. Postma and Vossen translated the same sets into Dutch [9]. Jin and Wu describe a shared task for Chinese semantic similarity, during which the WordSim353 set was translated [10].

Recently, the new similarity dataset Simlex-999 for English was created [5]. It contains 999 word pairs. In contrast to previous approaches, this dataset contains not only nouns but also verbs and adjectives. Besides, not only concrete words that have directly perceptible physical referents, but also abstract words were included. This dataset is constructed namely to reveal similarity (paradigmatic) relations between words. Currently, the Simlex-999 exists also in Italian, German, and Russian [6].

In 2015 the first Russian evaluation of word semantic similarity was organized [11][12]. One of the datasets created for the evaluation was the set of word pairs with the scores of their similarity obtained by human judgments (HJ-dataset). This dataset was prepared by translation of the existing English datasets MC, RG, and WordSim353.

Many approaches for detecting semantic word similarity are based on extracting word contexts within sentences [13], [14]. In this paper, we present a study of a possible contribution of features based on location of words in neighbor sentences. We conduct our experiments on the Russe HJ test set, the Russian translations the above-mentioned datasets: MC, RG and WordSim353 dataset divided into two parts: the similarity set and relatedness set.

The structure of the paper is as follows. In the second section we consider related work. The third section describes the Russe evaluation of Russian similarity tasks. In the fourth section we present our approach additionally utilizing neighbor sentence co-occurrence statistics of words. The fifth section describes the experiments and the achieved results.

II. Related work

The existing approaches for calculating semantic similarity between words are based on various resources and text collections. One of the well-known approaches utilizes the lexical relations described in manually created thesauri such as WordNet or Roget's thesaurus [4], [15] or such online resource as Wikipedia [16].

Another type of approaches for detecting semantic word similarity are distributional approaches that account for shared neighbors of words [14]. In these approaches a matrix, where each row represents a word w and each column corresponds

ISBN 978-952-68397-6-9 (paperback), ISBN 978-952-68397-7-6 (PDF)

to contexts of w, is constructed. The value of each matrix cell represents the association between the word w_i and the context c_j. A popular measure of this association is pointwise mutual information (PMI) and especially positive PMI (PPMI), in which all negative values of PMI are replaced by 0 [17], [18]. Similarity between two words can be calculated, for example, as the scalar product between their context vectors.

Recently, neural-network based approaches, in which words are “embedded” into a low-dimensional space, appeared and became to be widely used in lexical semantic tasks. Especially, word2vec tool, a program for creating word embedding, became popular [19]. In [17], [18] traditional distributional methods are compared with new neural-network approaches in several lexical tasks. Baroni et al. [17] found that the new approaches consistently outperform traditional approaches in most lexical tasks. For example, on the RG test traditional approaches achieve 0.74 Spearman’s correlation rank, the neural approach obtains 0.84.

Levy et al. [18] tried to find the best parameters for each approach and showed that on the WordSim353 subsets, the word2vec similarity is better on the WordSim353 similarity subset (0.773 vs. 0.755), while traditional distributional models achieve better results on the WordSim353 relatedness subset (0.688 vs. 0.623). They conclude that word2vec provides a robust baseline which "does not significantly underperform in any scenario".

Hill et al. [5] write about two important conclusions followed from the previous distributional model research:

- Models exploiting syntactic similarity better reveal similarity but approaches based on bag-of-words input better extract relatedness relations [4], [18];
- Models with larger context windows better reveal relatedness between words, small-context models better extract similarity relations [4].

Some authors suppose to combine various features to improve detection of word semantic similarity. Agirre et al. [4] use a supervised combination of WordNet-based similarity, distributional similarity, syntactic similarity, and context window similarity, and achieve 0.96 Spearman’s correlation on the RG set and 0.83 on the WordSim353 similarity subset.

The best result achieved on the Russe-2015 human judgment set [11], 0.7625 Spearman’s correlation, was obtained with a supervised approach combining word2vec similarity scores calculated on a united text collection (consisting of the ruwac web corpus, the lib.ru fiction collection, and Russian Wikipedia), synonym database, prefix dictionary, and orthographic similarity [20]. The second result (0.7187) was obtained with application of word2vec to two text corpora: Russian National Corpus and a news corpus. The word2vec scores from the news corpus were used if target words were absent in RNC [21]. The best result achieved by a traditional distributional approach was 0.7029 [11].

III. Russe Evaluation and Created Russian Datasets

The Russe shared task included three subtasks:

- HJ-subtask: detection of correlations with human judgments in terms of Spearman’s rank correlation,
- RT-subtask: classification of semantic lexical relations in terms of average precision, the automatic answers were compared with RuThes thesaurus relations [23],[32],
- AE-subtask: cognitive associations detection. The results of the participants were compared with two associative thesauri: Russian associative thesaurus [24] and Sociation thesaurus [25] .

The training data for all subtasks were provided. In case of the HJ subtask, a small development set consisting of 66 word pairs were given to the participants. The number of submissions from each participant was not restricted.

In this study we use HJ similarity datasets created for the Russe evaluation [11]. The data sets were constructed by translation of three English datasets: Miller-Charles set, Rubenstein-Goodenough set, and WordSim353. These datasets were joined together and translated into Russian in a consistent way. Each word of the datasets was translated with a single Russian noun.

The word similarity scores for this joint dataset were obtained by crowdsourcing from Russian native speakers. Each annotator was given an assignment consisting of 15 word pairs randomly selected from the Russe HJ set and has been asked to assess their similarity. The possible values of similarity were:

- 0 – not similar at all,
- 1 – weak similarity or relatedness,
- 2 – moderate similarity or relatedness,
- 3 – high similarity or relatedness.

By the end of the experiment, 4,200 answers were obtained that is 280 submissions of 15 judgment sets.

TABLE I. The first ten Russian word pairs from the Russian translation of WordSim353 similarity subset

N	Russian word pairs	English Counterparts
1	маг – волшебник (0.958)	magician – wizard (10)
2	машина – автомобиль (0.952)	car – automobile (13)
3	мальчик – парень (0.952)	boy – lad (16)
4	доллар – бакс (0.952)	dollar – buck (5)
5	расчет – вычисление (0.917)	calculation – computation (24)
6	побережье – берег (0.905)	shore – coast (7)
7	жидкость – вода (0.905)	liquid–water(36)
8	чемпионат – турнир (0.889)	championship – tournament (26)
9	тигр – тигр (0.875)	tiger – tiger (1)
10	поездка – путешествие (0.857)	voyage – journey(3)

All estimated word pairs were subdivided to the training set (66 word pairs) for adapting the participants’ methods and the test set, which was used for evaluation (335 word pairs). Correlations with human judgments were calculated in terms of Spearman's rank correlation. Currently, the full Russian similarity dataset is published[12].

TABLE II. THE FIRST TEN RUSSIAN WORD PAIRS FROM THE RUSSIAN TRANSLATION OF WORDSIM353 RELATEDNESS SUBSET.

N	Russian word pairs	English Counterpart
1	расположение – размещение (0.917)	arrangement – accommodation (137)
2	деньги – богатство (0.852)	money – wealth (12)
3	ребенок – мать (0.815)	mother – child (29)
4	теннис – ракетка (0.7)	tennis – racket (41)
5	новость – сообщение (0.667)	news – report (14)
6	война – войска (0.667)	war – troops (15)
7	книга – библиотека (0.667)	book – library (48)
8	банк – деньги (0.636)	bank – money (17)
9	морепродукт – море (0.636)	seafood – sea (47)
10	деньги – банк (0.625)	money – bank (4)

In this study we present our results on the Russe HJ test dataset (335 word pairs). Besides, we singled out the Russian translations of MC and RG similarity sets, WordSimS353 similarity and WordSim353 relatedness sets from the Russe HJ similarity dataset and show our results on each dataset.

Table I presents the first ten most similar word pairs from the Russian translation of WordSim353 similarity set according to Russian native speakers and gives the comparison with its English counterparts. Table II depicts the most related words for the Russian translation of WordSim353 relatedness subset. The number in parentheses for English counterparts indicates the rank of this pair in the corresponding English testsets.

IV. USING NEIGHBOR SENTENCES FOR WORD SIMILARITY TASK

Analyzing the variety of features used for word semantic similarity detection, we could not find any work utilizing co-occurrence of words in neighbor sentences. However, repetitions of semantically related words in natural language texts bear a very important role providing the cohesion of text fragments [26], that is a device for "sticking together" different parts of the text. Cohesion can be based on the use of semantically related words, reference, ellipsis, or conjunctions. Among these types, the most frequent type is lexical cohesion, which is usually expressed with sense-related words such as repetitions, synonyms, hyponyms, etc.

When analyzing a specific text, the location of words in sentences can be employed for constructing lexical chains. In classical works [27], [28], the location of words in neighbor sentences is used for adding new elements in the current lexical chain. The distance in sentences between referential candidates is an important feature in such tasks as anaphora resolution and coreference chains construction [29]. In [30] the authors use the frequency of word co-occurrence in neighbor sentences as an additional feature to improve text summarization.

Thus, in a coherent text, following sentences are often lexically attached to previous sentences. It allows us to suppose that the analysis of word distribution in neighbor sentences is able to give additional information about their semantic similarity. If two words often co-occur near each other it often indicates that they are components of the same collocation. Frequent co-occurrence of two words in the same sentences possibly means that they correspond to participants of the same situations [13]. But if words often co-occur in neighbor sentences it could show that they are semantically or thematically similar.

In this study, we calculate co-occurrence of words in two sentences that are direct neighbors to each other and analyze its importance for lexical similarity detection in form of two basic features.

The first feature based on co-occurrence of words in neighbor sentences (NS) is calculated if words are located in different sentences in the following way: the first word w_1 should be mentioned only in the first sentence and the second word w_2 should occur only in the second sentence. If a word occurs in both sentences we do not calculate the NS feature in this pair of sentences for this word because we cannot distinguish the factor of the neighbor sentence co-occurrence from the same sentence co-occurrence.

The second feature (TS) is the ordinary frequency of word co-occurrence in a large window equal to two sentences.

The following example from English Wikipedia demonstrates the difference in calculation of TS and NS features:

Over time, the word ***"automobile"*** *fell out of favour in Britain, and was replaced by "motor* ***car****". An abbreviated form, "auto", was formerly a common way to refer to* ***cars*** *in English, but is now considered old-fashioned.*

In this example, the TS feature is equal to 2 because we see two lemmas "car" in the large two-sentence window. But the NS feature is equal to 0 because lemma "car" is repeated in both sentences and this context is not considered as informative from the point of view of the lexical cohesion phenomenon.

We consider these features in two variant weightings: pointwise mutual information (pmiNS, pmiTS) and normalized pointwise mutual information (npmiNS, npmiTS). It should be also mentioned that both types of features do not require any tuning because they are based on sentence boundaries. It means that they are self-adaptive to a text collection.

To compare these features with features extracted from a window within the same sentences, we calculated the feature IS (inside-sentence window), with its variants pmiIS, npmiIS. These features show co-occurrence of a word pair in a specific word window within a sentence. We experimented with various sizes of word windows and present only the best results usually based on a large window of 20 words (10 words to the left and 10 words to the right within a sentence). In fact, in most cases IS-window is equal to the whole sentence. Besides, we calculated a pointwise normalized mutual information for word pair co-occurrences within documents (npmiDoc).

We calculate PMI and NPMI using formulas 1 and 2. For calculating npmiDoc, the frequency of word co-occurrence in documents and document frequencies are employed. In other cases word co-occurrences in text fragments and collection frequencies are calculated.

Thus, NPMI is a normalized version of PMI [31]. It is bounded in the [-1,1] segment and has a transparent interpretation. Word pairs with negative NPMI co-occur rarer than independent events, positive NPMI means more frequent co-occurrence than expected. The maximum value of NPMI implies that both words always co-occur. NPMI=0 means independence of word co-occurrences.

$$pmi(w_1, w_2) = log\frac{p(w_1, w_2)}{p(w_1)p(w_2)} \qquad (1)$$

$$npmi(w_1, w_2) = log\frac{pmi}{-log(p(w_1, w_2))} \qquad (2)$$

where $p(w_1, w_2)$ is the probability to meet a word pair in a specific text fragment, $p(w_1), p(w_2)$ - probabilities of words in a text collection or probabilities to find a document, containing a word (for npmiDoc).

TABLE III. The results obtained on the Russe test set. The best single features and pairs of features are highlighted

Features	News Collection	Ru-Wiki	United Collection
NS	0.519	0.494	0.543
pmiNS	0.632	0.598	0.660
npmiNS	0.638	0.612	0.677
TS	0.604	0.580	0.607
pmiTS	0.645	0.627	0.657
npmiTS	0.648	0.636	0.670
IS	0.577	0.601	0.594
pmiIS	0.645	0.633	0.652
npmiIS	**0.663**	**0.658**	**0.681**
npmiDOC	0.611	0.567	0.641
best of w2v	0.651	0.618	0.680
NS+w2v	0.659	0.633	0.674
pmiNS+w2v	0.687	0.647	**0.712**
npmiNS+w2v	0.661	0.652	0.704
TS+w2v	0.679	0.662	0.689
pmiTS+w2v	**0.694**	0.663	0.710
npmiTS+w2v	0.679	0.662	0.697
IS+w2v	0.657	0.646	0.679
pmiIS+w2v	0.693	0.670	0.710
npmiIS+w2v	0.689	**0.674**	0.705
npmiDoc+w2v	0.659	0.619	0.692
npmiNS+npmiIS	0.665	0.652	0.689
npmiTS+npmiIS	0.661	0.652	0.676

To compare with state-of-the art approaches, we utilized the word2vec tool for processing our collections [19]. For preprocessing, the text collections were lemmatized, all function words were removed and, thus, the word2vec similarity(w2v) between words was calculated only on the contexts containing nouns, verbs, adjectives, and adverbs. The word2vec similarity was calculated with various parameters and only the best results are presented in the paper. The number of dimensions is in most cases equal to 300 but the window sizes can be quite different in specific cases. The following window sizes (1, 2, 3, 4, 5, 10, 15, 20) were tested.

We also considered simple combinations between pairs of features calculated as summation of the ranks of word pairs in the ordering determined with those features.

V. Experiments

We experimented with the following types of datasets: Russe HJ test set, MC and RG sets, which aim at distinguishing paradigmatic relations from unrelated pairs, and both Wordsim353 sets (WordSim353 similarity and relatedness sets). We use two texts collections: Russian Wikipedia (0.24 B tokens) and Russian news collection (0.45 B tokens). Besides, we also experimented on the united collection containing both the above-mentioned collections. The comparison of obtained automatic results with human judgements is calculated using Spearman's correlation rank.

TABLE IV. The results obtained on the Russian translation of the MC set. The best single features and pairs of features are highlighted

Features	News Collection	Ru-Wiki	United Collection
NS	0.746	0.611	0.791
pmiNS	0.728	0.790	0.735
npmiNS	0.803	**0.833**	0.827
TS	0.795	0.691	0.787
pmiTS	0.676	0.797	0.706
npmiTS	**0.834**	0.832	0.823
IS	0.703	0.710	0.770
pmiIS	0.700	0.784	0.742
npmiIS	0.805	0.824	0.817
npmiDoc	0.654	0.737	0.687
best of w2v	0.812	**0.833**	**0.843**
NS+w2v	0.841	0.785	0.848
pmiNS+w2v	0.821	0.846	0.836
npmiNS+w2v	0.837	**0.861**	0.868
TS+w2v	**0.872**	0.816	**0.881**
pmiTS+w2v	0.820	0.844	0.837
npmiTS+w2v	0.858	0.850	0.863
IS+w2v	0.838	0.812	0.879
pmiIS+w2v	0.834	0.832	0.843
npmiIS+w2v	0.852	0.843	0.863
npmiDoc+w2v	0.788	0.816	0.828
npmiNS+npmiIS	0.810	0.840	0.821
npmiTS+npmIS	0.814	0.834	0.825

Table III presents the results for the Russe HJ test set on three collections: the news collection, Ru-wiki, and the united collection. The best single features and pairs of features are highlighted. The best w2v window sizes for all collections are equal to 5.

It can be seen that among single features the word2vec ranking is always better than simple frequencies of co-occurrences for all measures (NS, IS, TS). It is also better than PMI weighting in most cases excluding Wikipedia-based measures. But NPMI-based weighting of IS-feature exceeds word2vec on all three collections.

If to consider the pair of features, one can see that all combinations of PMI and NPMI-based word co-occurrences with w2v improve the results considerably. The best achieved results on two collections (the news collection and the united collections) are based on combination of the w2v feature and two different two-sentence based features. This result is very close to the second result achieved during the Russe evaluation (0.717) but in that approach, a three times larger news collection and balanced Russian National Corpus (RNC) were used [21].

Thus, on the Russe test set, the best result was obtained with additional accounting for co-occurrence of words in neighbor sentences. Though it should be noted that combination of word2vec with an inside-sentence feature generated a very similar result (0.710).

The specificity of the Russe test set is that it contains paradigmatic relations (proper semantic similarity) together

TABLE V. THE RESULTS OBTAINED ON THE RUSSIAN RG SET. THE BEST SINGLE FEATURES AND PAIRS OF FEATURES ARE HIGHLIGHTED

Features	News Collection	Ru-Wiki	United Collection
NS	0.696	0.568	0.730
pmiNS	0.711	0.695	0.747
npmiNS	0.760	0.750	0.813
TS	0.754	0.704	0.778
pmiTS	0.685	0.762	0.737
npmiTS	0.779	0.806	0.820
IS	0.693	0.726	0.772
pmiIS	0.688	0.762	0.742
npmiIS	0.744	**0.807**	0.811
npmiDoc	0.640	0.720	0.737
best of w2v	**0.791**	0.805	**0.858**
NS+w2v	0.833	0.795	0.860
pmiNS+w2v	0.819	0.827	0.847
npmiNS+w2v	0.835	0.850	0.868
TS+w2v	**0.841**	0.842	**0.870**
pmiTS+w2v	0.808	0.835	0.843
npmiTS+w2v	0.831	**0.858**	0.863
IS+w2v	0.825	0.838	0.865
pmiIS+w2v	0.814	0.834	0.845
npmiIS+w2v	0.826	0.851	0.857
npmiDoc+w2v	0.761	0.798	0.842
npmiNS+npmiIS	0.781	0.811	0.820
npmiTS+npmiIS	0.765	0.812	0.811

TABLE VI. THE RESULTS OBTAINED ON THE RUSSIAN TRANSLATION OF WORDSIM353 SIMILARITY SET

Features	News Collection	Ru-Wiki	United Collection
NS	0.513	0.579	0.562
pmiNS	0.617	0.656	0.675
npmiNS	0.639	0.693	0.702
TS	0.626	0.653	0.646
pmiTS	0.648	0.670	0.676
npmiTS	0.661	0.690	0.693
IS	0.591	0.666	0.613
pmiIS	0.638	0.671	0.665
npmiIS	0.660	**0.708**	0.690
npmiDoc	0.653	0.608	0.690
best of w2v	**0.727**	**0.708**	**0.747**
NS+w2v	0.691	0.721	0.737
pmiNS+w2v	0.722	0.735	**0.760**
npmiNS+w2v	0.719	0.740	0.758
TS+w2v	0.725	0.741	0.746
pmiTS+w2v	**0.741**	0.674	**0.760**
npmiTS+w2v	0.728	0.738	0.752
IS+w2v	0.711	0.744	0.733
pmiIS+w2v	0.731	0.742	0.746
npmiIS+w2v	0.726	**0.747**	0.747
npmiDoc+w2v	0.733	0.707	0.752
npmiNS+npmiIS	0.666	0.711	0.706
npmiTS+npmiIS	0.665	0.705	0.697

with relatedness relations from the translated the WordSim353 set. Therefore, we decided to study the proposed features on the Russian translations of RG and MC sets, which are constructed for measuring paradigmatic relations between words. Tables IV and V present the results obtained for these datasets. The best w2v window sizes for these calculations are in the interval (1–3). The presented results are the first results obtained for these Russian datasets.

We can see quite a different picture on these sets than on the Russe test set. The positions of w2v similarity considerably improved on all text collections. In both cases the maximal values of Spearman's correlation rank for all three collections are obtained with the combination of word2vec and features based on word co-occurrence in neighbor sentences. In our opinion, it means that the frequent co-occurrence of words in neighbor sentences bears additional useful information for detecting paradigmatic lexical similarity.

To experiment with similarity and relatedness separately, we repeated the same evaluations on the Russian translations of WordSim353 similarity and relatedness datasets singling them out from the whole Russian word pairs translated in the Russe framework. The results of the best features and pairs of features are shown in Tables VI-VII. The best features and feature combinations for each set are highlighted. The best w2v window sizes for the Russian WordSim353 similarity set is equal to 1 on all collections. For the Russian WordSim353 relatedness set the best windows are also small: 1–3 words.

The Table VIII shows the first ten related word pairs from Russian WordSim353 relatedness set according to the best feature pair and word2vec similarity. In parentheses the rank of a word pair in the corresponding human judgement list is indicated.

We can conclude from the Tables III-VII:

- a simple combination of two features (rank summation) leads to considerable improvement of the results achieved by single features. The combination of npmiTS and word2vec features was always better than the best single feature for all datasets and text collections, even for Ru-Wiki;
- the behavior of features is quite different on the news collection and the Ru-Wiki, the significance of IS-based features (inside-sentence co-occurrence) is quite larger for Ru-Wiki. It can be explained with the explanatory character of Wikipedia articles.

The obtained results are the first ones for the Russian translations of the WS353 similarity and relatedness sets.

TABLE VII. THE RESULTS OBTAINED ON THE RUSSIAN WORDSIM353 RELATEDNESS SET

Features	News Collection	Ru-Wiki	United Collection
NS	0.549	0.499	0.567
pmiNS	0.641	0.641	0.661
npmiNS	**0.671**	0.659	0.691
TS	0.585	0.569	0.602
pmiTS	0.643	0.661	0.665
npmiTS	0.663	0.675	0.687
IS	0.549	0.593	0.573
pmiIS	0.629	0.665	0.655
npmiIS	0.656	**0.689**	0.685
npmiDoc	0.609	0.583	0.630
best of w2v	**0.671**	0.636	**0.701**
NS+w2v	0.667	0.635	0.701
pmiNS+w2v	0.702	0.673	0.719
npmiNS+w2v	0.703	0.678	**0.728**
TS+w2v	0.683	0.662	0.713
pmiTS+w2v	**0.706**	0.684	0.722
npmiTS+w2v	0.704	0.691	0.727
IS+w2v	0.671	0.677	0.699
pmiIS+w2v	0.700	0.693	0.717
npmiIS+w2v	0.701	**0.700**	0.725
npmiDoc+w2v	0.680	0.645	0.705
npmiNS+npmiIS	0.678	0.691	0.700
npmiTS+npmiIS	0.665	0.688	0.690

VI. CONCLUSION

In this paper we presented the first results of detecting word semantic similarity on the Russian translations of several semantic word similarity datasets, including Miller-Charles

TABLE VIII. BEST APPROACH AND BEST WORD2VEC APPROACH FOR THE WORDSIM3535 RELATEDNESS SETS

N	Most related words according to the best feature pair	Most related words according to the best word2vec
1	чашка – кофе (27) (cup–coffee)	команда – игра (70) (command – team)
2	опек – нефть (77) (OPEC–oil)	год – начало (112) (year– start)
3	иерусалим – израиль (13) (Jerusalem– Israel)	израиль – иерусалим (13) (Jerusalem– Israel)
4	иерусалим – палестинец (39) (Jerusalem – Palestinian)	кофе – чашка (27) (cup – coffee)
5	ливень – наводнение (28) (shower – flood)	наводнение – ливень (28) (shower – flood)
6	теннис – ракетка (4) (tennis – racket)	театр – кино (19) (theater – movie)
7	книга – библиотека (7) (book – library)	палестинец – иерусалим (39) (Palestian –Jerusalem)
8	планета – астроном (44) (planet – astronomer)	мать – ребенок (3) (mother–child)
9	игра – команда (70) (game – team)	экология – среда (68) (ecology – environment)
10	компьютер – клавиатура (20) (computer – keyboard)	компьютер – интернет (34) (computer – Internet)

and Rubenstein-Goodenough sets, WordSim353 similarity and relatedness sets, prepared for the first Russian word semantic evaluation Russe-2015. The experiments were carried out on three text collections. We studied the contribution of two-sentence co-occurrence to detect the similarity and relatedness between words.

We found that the best results in detection of similarity and relatedness semantic relations are achieved using the combination of word2vec and the co-occurrence of words in neighbor sentences. A simple combination of word2vec similarity and npmiTS features is better than the best single feature for all datasets and text collections under consideration. Besides, we found that the behavior of features is different on the Wikipedia collection, for which the best results are correlated with npmi weighting of inside-sentence co-occurrence of words.

In future we plan to add information from the existing Russian thesauri such as RuThes [22] or RuWordNet [32] to the process of word similarity or relatedness detection. The use of additional knowledge will be based on an unsupervised approach exploiting properties of the natural language text.

Besides, we think that the important role of co-occurrence in neighbor sentences can be used in statistical topic modeling [33].

ACKNOWLEDGMENT

This work is partially supported by the Russian Foundation for Basic research (grant N14-07-00383) and Russian Foundations for Humanities (grant N15-04-12017).

REFERENCES

[1] H. Rubenstein and J.B. Goodenough, "Contextual Correlates of Synonymy", *Communications of the ACM*, Vol. 8(10), 1965, pp. 627-633.

[2] G. Miller and W.G. Charles, "Contextual correlates of semantic similarity", *Language and Cognitive Processes*, 6(1), 1991, pp. 1-28.

[3] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman and E. Ruppin, "Placing Search in Context: The Concept Revisited", *in. Proc. of the 10th International Conference on World Wide Web WWW-2001*, Hong Kong, China, 2001, pp.406-414.

[4] E. Agirre, E. Alfonseca, K. Hall, J, Kravalova, M. Paşca, M. and A. Soroa,"A study on similarity and relatedness using distributional and wordnet-based approaches", *in Proc. the 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics NAACL-2009*, 2009, pp.19-27.

[5] F. Hill, R. Reichart and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation", *Computational Linguistics*, 4, 2015, pp. 665-695.

[6] I. Leviant and R. Reichart, "Judgment Language Matters: Multilingual Vector Space Models for Judgment Language Aware Lexical Semantics", *arXiv preprint* arXiv:1508.00106, 2015.

[7] I. Gurevych, "Using the Structure of a Conceptual Network in Computing Semantic Relatedness", in *Proc. of the 2nd International Joint Conference on Natural Language Processing*, 2005, pp. 767-778.

[8] S. Hassan and R. Mihalcea, "Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge", *in Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing EMNLP-2009*, Vol. 3, Singapore, 2009, pp.1192-1201.

[9] M.Postma and P. Vossen, "What implementation and translation teach us: the case of semantic similarity measures in wordnets", in Proc. of Global WordNet Conference GWC-2014, Tartu, Estonia, 2014, pp.133-141.

[10] P. Jin and Y. Wu, "Semeval-2012 task 4: evaluating Chinese similarity", *in Proc. of the First Joint Conference on Lexical and Computational Semantics Volume 1: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 2012, pp. 374-377.

[11] A. Panchenko, N. Loukachevitch, D. Ustalov, D. Paperno, C. Meyer and N. Konstantinova, "Russe: The first workshop on Russian semantic similarity", *in Proc. Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference. Dialogue-2015*, Vol. 2, 2015, pp.89-105.

[12] The First International Workshop on Russian Semantic Similarity Evaluation (RUSSE), Web: http//russe.nlpub.ru/.

[13] M. Sahlgren, "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high dimensional vector spaces", Ph.D. thesis, Univ. of Stockolm, 2006.

[14] G. Lapesa and S. Evert, "A large scale evaluation of distributional semantic models: Parameters, interactions and model selection", *Trans. of the Association for Computational Linguistics*, 2, 2014, pp. 531-545.

[15] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness", *Computational Linguistics*, 32(1), 2006, pp. 13-47.

[16] E.Gabrilovich and S. Markovitch, "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis", *in Proc. of Jount Intern. Conf. on Artificial Intelligence IJCAI-2007*, 2007, pp. 6-12.

[17] M.Baroni, G. Dinu and G. Kruszewski, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors", *in Proc. of Annual Conference of Association on Computational Linguistics ACL-2014*, 2014, pp.238-247.

[18] O. Levy, Y. Goldberg and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings", *Transactions of the Association for Computational Linguistics*, 3, 2015, pp. 211-225.

[19] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", *in Proc. Advances in Neural Information Processing Systems NIPS-2013*, 2013, pp.3111–3119.

[20] K.A. Lopukhin, A.A. Lopukhina and G.V. Nosyrev, "The impact of different vector space models and supplementary techniques on Russian semantic similarity task", *in Proc. Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference Dialogue-2015*, Vol. 2, 2015, 145-153

[21] A Kutuzov and E. Kuzmenko, "Comparing Neural Lexical Models of a Classic National Corpus and a Web Corpus: The Case for Russian", *in Proc. Computational Linguistics and Intelligent Text Processing Cicling-2015*, Springer International Publishing, 2015, pp.47-58.

[22] N. Loukachevitch and B. Dobrov, "RuThes Linguistic Ontology vs. Russian Wordnets", *in Proc. of Global WordNet Conference GWC-2014*, Tartu, 2014, pp. 154-162.

[23] RuThes Thesaurus, Web: http://www.labinform.ru/pub/ruthes/

[24] Russian associative thesaurus, Web: http://it-claim.ru/asis.

[25] Sociation thesaurus, Web: http://sociation.org.

[26] M. Halliday and R. Hasan, *Cohesion in English*, Routledge, 1976.

[27] R.Barzilay and M. Elhadad, “Using lexical chains for text summarization”, *Advances in automatic text summarization*, 1999, pp. 111-121.

[28] G. Hirst, G. and D. St-Onge, “Lexical chains as representations of context for the detection and correction of malapropisms”, *WordNet: An electronic lexical database*, 1998, pp.305-332.

[29] E.R. Fernandes, C.R. dos Santos and R.L.Milidiú, “Latent trees for coreference resolution”, *Computational Linguistics*, 2014.

[30] N. Loukachevitch and A. Alekseev, “Summarizing News Clusters on the Basis of Thematic Chains”, *in Proc. Language resources and evaluation conference LREC-2014*, 2014, pp. 1600-1607.

[31] G. Bouma, “Normalized (pointwise) mutual information in collocation extraction”, *in Proc. of German Society for Computational Linguistics Conference GSCL-2009*, 2009, pp. 31-40.

[32] N. Loukachevitch, G. Lashevich, A. Gerasimova, V. Ivanov and B. Dobrov, “Creating Russian Wordnet by Conversion”, *in Proc. of Computational Linguistics and Intellectual Technologies Conference Dialog-2016*, 2016.

[33] D.M.Blei, A.Y. Ng and M.I. Jordan, “Latent dirichlet allocation”, *the Journal of Machine Learning Research*, 3,2003, pp. 993-1022.