Comparison of Sentence Similarity Measures for Russian Paraphrase Identification

Ekaterina Pronoza, Elena Yagunova
Saint-Petersburg State University
Saint-Petersburg, Russian Federattion
{katpronoza, iagounova.elena}@gmail.com

Abstract—**In this paper we analyze and compare different types of sentence similarity measures applied to the problem of sentential paraphrase identification. We work with Russian, and all the experiments are conducted on the Russian paraphrase corpus we have collected from the news headlines (and are collecting at the moment). Apart from the similarity measures, we also analyze the corpus itself. As a result of the research we disprove the supposition that it is more difficult to distinguish between precise and loose paraphrases than between loose paraphrases and non-paraphrases. We also come up with the recommendations for the application of different similarity measures to identifying paraphrases derived from the news texts.**

I. Introduction

This paper describes our work on the ongoing project ParaPhraser.ru [16]. The project is dedicated to paraphrase extraction, identification and generation. As part of the project, we have already collected a corpus of 6281 sentence pairs annotated as precise (1482), loose (3247) and non-paraphrases (2209), and it is constantly increasing in size. The corpus is annotated via crowdsourcing at http://paraphraser.ru.

Although we distinguish three classes of paraphrases, precise and loose paraphrases are actually of the main interest to us during corpus construction and in further work: to be able to use the corpus in natural language processing tasks like information extraction and text summarization (as we intend to do) we need to distinguish these two types of paraphrases. Hence, negative instances are only kept for contrast purposes. They include not only paraphrase candidates approved by the unsupervised similarity metric (used for corpus construction) and rejected by the annotators but also a small portion of random sentence pairs rejected by both the similarity metric and the annotators. The latter sentences are added to mitigate paraphrase classes imbalance.

Thus, we initially believed that the most difficult task for both the annotators and the paraphrase identification model trained on the corpus would be to distinguish between precise and loose paraphrases because the difference between them can be very subtle. In this paper this supposition is checked and disproved via the series of experiments and the analysis of annotated sentence pairs.

We experiment with three different types of sentence similarity measures: shallow measures, semantic dictionary-based measures and distributional semantic measures. Shallow measures mostly correspond to the overlap in the sentences on the phrase, word and character levels. They usually include metrics like BLEU, longest common substring (LCS), skip-grams overlap and others, proposed in the earlier papers on paraphrase identification and sentence similarity. Dictionary-based semantic measures employ external semantic resources to predict the similarity of the sentences. For example, for English such measures usually use WordNet. For Russian there is YARN, a large open WordNet-like machine-readable thesaurus created via crowdsourcing [3]. Apart from it, we also use the dictionary of word formation families [28]. And finally, distributional semantic measures are based on vector space models (or distributional semantic models).

We analyze the characteristics of sentences on which the outlined types of measures fail. Both the characteristics of the similarity measures and the paraphrase corpus are revealed. The results of the misclassification analysis give us the intuition of the possible improvements of both the paraphrase identification model and the annotation of the corpus. We also hope that our results will help other researchers who work with Russian (and similar languages with free word order and rich morphology) in developing their paraphrase identification models.

II. Related Work

Based on the previous research on sentence similarity measures and paraphrase identification, we can classify existing sentence similarity measures into the following groups:

1) Shallow (based on string or lexical overlap).
2) Semantic (based on the semantic structure of the sentences, using external semantic resources).
3) Syntactic (based on the syntactic structure of the sentences).
4) Distributional (based on vector space models, or distributional semantic models).

Shallow similarity measures are the earliest similarity measures used in paraphrase and semantic communities. They are mainly based on the overlap of words, phrases or characters [3], [10], [11]. We also classify metrics originating from machine translation, like BLEU [17], [25], [27], as shallow, if they are based on the surface forms of the words and do not employ any semantic resources. Other shallow features include edit distance between the sentences [3], [17], [27], sentence length difference [11], [25], the length of the longest

ISBN 978-952-68397-0-7 (paperback), ISBN 978-952-68397-1-4 (PDF)

common subsequence [6], [13], [17], [27], the number of matching proper names and cardinal numbers [6], [13], etc.

Most dictionary-based semantic similarity measures use WordNet [15] or WordNet-like resources and exploit synonymy or hypernymy relations. Roughly speaking, the similarity between the words can be calculated as the length of the shortest path between them in the WordNet graph. A comprehensive study of different WordNet-based measures can be found in [4]. Such measures are more sophisticated than the shallow ones, but they have evident limitations because they are strongly dependent on the quality and coverage of the corresponding semantic resources.

Another approach (syntactic measures) is applied in [8], [20], [22], etc. and often implies the use of dependency pairs (dependency relations overlap calculated as precision and recall, edit distances between syntactic parse trees, etc.). Sidorov et al. [23] propose to calculate the similarity between texts as the similarity between their respective syntactic n-grams using tree edit distance. Such measures allow us to capture even more linguistic phenomena than the previously described ones but as they use the output of a syntactic parser, they propagate errors made by the parser.

Distributional semantic measures can serve as an alternative to both semantic and syntactic measures as they can predict semantic similarity without analyzing the deep structure of the sentences. For example, in [1] it is shown that on high-complexity datasets like Microsoft Research Paraphrase Corpus (MSRP) [3] and the third recognising textual entailment challenge (RTE3) dataset [7] overlap-based and distributional measures perform better than the linguistic (semantic and syntactic) ones. The distributional approach is based on the supposition that semantically close words occur in similar contexts. Distributional models can be classified into count-based (e.g., based on Latent Semantic Analysis (LSA)) and predictive ones (e.g., skip-grams and bag-of-words models implemented in Word2vec [14] [14]). According to a recent research [1], predictive models outperform count-based ones on a wide range of semantic tasks (semantic relatedness, synonym detection, concept categorization, etc.).

In paraphrase identification community, state-of-the-art results on MSRP are usually achieved by using a combination of different types of similarity measures and/or by tuning distributional measures, e.g., by using discriminative TF-IDF weighting (TF-KLD) [11] together with fine-grained linguistic features or discriminative TF-IDF weighting for both words and phrases with smoothing based on the K nearest neighbours (KNN) algorithm (TF-KLD-KNN) [26]. In this paper we focus on the analysis of different types of similarity measures, and tuning them is beyond our task at the moment. Instead, our aim is to provide the comparison of similarity measures with respect to their performance on different types of sentence pairs.

A study of different similarity measures was already conducted, for example, in [1]. The authors considered 3 types of measures: overlap-based measures (i.e., shallow), linguistic measures and the measures based on vector space model with TF-IDF weighting and cosine distance. These 3 measures were tested against 3 different datasets for English with 2 paraphrase classes: paraphrases and non-paraphrases. Unlike [1], we consider a substantially extended set of shallow features, and over 40 different vector distances other than the cosine distance. In vector space model we also adopt a bag-of-words approach but do not apply any weighting. We believe that the preliminary analysis should be conducted to select the appropriate weighting scheme. We also distinguish semantic and syntactic linguistic similarity measures (and focus on the semantic ones in this paper). As we work with Russian, which is a morphologically rich language, our semantic similarity measures are also extended: they combine the information about synonymy relations and word formation families.

III. DATA

Our paraphrase corpus consists of Russian sentence pairs collected from news headlines. Several Russian media sources are parsed every day in real time, their headlines are compared using an unsupervised similarity metric described in [16] and candidate pairs are included in the corpus. They are further evaluated via crowdsourcing and labeled as precise, loose or non-paraphrases. At the moment there are 6281 sentence pairs (1482 precise, 3247 loose and 2209 non-paraphrases). Most negative instances are represented by the unsuccessful paraphrase candidates, but some, although not being paraphrase candidates according to the unsupervised similarity metric, are still added to the corpus to make it more balanced. Both types of negative instances are considered non-paraphrases when rejected by the annotators (i.e., labeled as non-paraphrases).

It should be mentioned that the corpus is expected to have high degree of word overlap in the paraphrases. There are several reasons for that. Firstly, the language of news reports does not vary much. News headlines, from which the corpus is collected, are even more laconic and their style is similar in different media sources. Secondly, pairs of headlines are included into the corpus based on the values of the similarity metric which incorporates both semantic similarity and string-level similarity between the sentences. This metric extends the one proposed in [10] and can also be called a variant of soft cosine similarity measure [22]. For a general-purpose paraphrase corpus, high word overlap could be a serious drawback, but in our case the corpus is created for the use in information extraction and text summarization where the data is often represented by the news texts.

IV. SENTENCE SIMILARITY MEASURES

In this section we describe the three types of similarity measures we experiment with.

A. Shallow Measures

1) Traditional measures: Measures traditionally used in paraphrase identification.

These measures include words/characters overlap, sentence length difference, edit distance, longest common subsequence, BLEU and others described in Section II. In our experiments they are represented by 13 features.

2) Our measures: String-, word- and lemma-level measures based on the differences between the sentences.

Such measures are described in detail in our paper in press [18]. They are calculated based on the characteristics of what is left in the two sentences after the removal of overlapping words (e.g., the portion of notional/capitalized words, the portion of overlapping substrings left in the two sentences after such procedure, etc.).

B. Dictionary-based Semantic Measures

Most semantic similarity measures for English are based on WordNet. For Russian we employ two resources: YARN – Yet Another RussNet [3] and the dictionary of word formation by Tikhonov [28]. The former includes several thousands of synsets and is a rich source of synonyms and the latter contains information about families of words in Russian, i.e., words with the same root (which are also semantically related). Such information is quite important for detecting sentence similarity in Russian, with its rich morphology.

In our experiments we work with the improved version of the similarity metric we used for corpus construction (which is actually a semantic measure). Other semantic measures are based on the similarities between the sentences with removed overlapping words: namely, on the portion of synonyms, words with the same root, synonyms of the synonyms, synonyms of the words with the same root, etc. The details of their calculation can also be found in [18].

C. Distributional Semantic Measures

In most papers in NLP the use of distributional semantic models implies the calculation of cosine distance; however, cosine distance is not a panacea. We experiment with 44 different vector distance measures described in a comprehensive study of distance/similarity measures between density probability functions [6]. The authors describe a wide range of different measures which can be used for vectors comparison. We employ all of these measures in our research with the exception of those which can be represented as the linear combination of the other measures. Some of the measures are strongly correlated, some are not. We do not try to separate such measures and to select one measure from each correlating group. Instead, we leave this task to the classifier.

Apart from the 44 vector distance measures, we calculate cosine distance (separately). We also use 2 variations of cosine distance measure: (1) firstly, we calculate it only on the words the sentences differ in and (2) secondly, we use cosine distance as an inner matrix similarity measure in the improved unsupervised similarity metric (which was initially used for corpus construction) [18].

It should be noted that in all our distributional measures vector distances are calculated between sentence vectors, and sentence vectors are calculated as the average of their words vectors. Thus, we adopt a bag-of-words approach at the moment and do not apply any vector weighting. The vectors of words are calculated according to the skip-gram model implemented in Word2vec.

V. Experiments

A. Tools

In our experiments the described similarity measures are used as features in the paraphrase classification task. To calculate feature values which involve lemmatization and POS-tagging, we use TreeTagger [21]. For distributional semantic features Word2vec skip-gram model is trained on the news corpora from 4 different sources. These corpora consist of about 4.3 million sentences, 65.8 million tokens and contain news reports from 2012 and 2013 (which are not included in our paraphrase corpus). The context window size is 7 words (i.e., (−7, +7)), frequency threshold equals 4, and the dimensionality of feature vectors is set to 300.

All our experiments, including classifiers training, feature selection and parameters tuning, are conducted using scikit-learn and pandas.

B. Similarity Measures

In this paper we experiment with several types of features:

1) shallow (based on n-gram/word/substring/character overlap, word shape (capitalization));
2) POS (based on the overlap of words of particular part of speech);
3) semantic, dictionary-based (using synonymy relations, word families information; semantic information is derived from the dictionaries);
4) semantic, distributional (using distributional semantic models; semantic information derived from the corpora).

We initially experimented with 4 types of features and their combinations, but as POS features did not contribute to the overall result at all, we eliminated them from any further experiments and analysis.

We have the following feature sets:

1) shallow (13 traditional and 11 newly introduced features);
2) POS (5 features);
3) semantic (stands for dictionary-based features) – 11 features;
4) distrib (stands for distributional semantic) – 44 features;
5) cosine (we experiment with cosine distance separately);
6) extended cosine (extended cosine distance measure: 2 additional variations of cosine distance);

and their combinations:

7) shallow + POS;
8) distrib* + cosine;
9) distrib + cosine + extended cosine;
10) shallow + semantic;
11) shallow + semantic + distrib;
12) shallow + semantic + cosine;
13) shallow + semantic + cosine + extended cosine;
14) shallow + semantic + distrib + cosine;
15) shallow + semantic + distrib + cosine + extended cosine;
16) shallow + POS + semantic + distrib + cosine + extended cosine.

Distrib stands for features calculated using over 40 various vector similarity measures [6] (cosine similarity not included).

We evaluate the features separately and in various combinations and analyze their performance against our paraphrase corpus.

The features are used with GradientBoostingClassifier. One of the recent directions for the improvement of standard classifiers involves ensembles of classifiers and we decided to adopt such an approach. We initially experimented with AdaBoost, Random Forests and Gradient Tree Boosting (GTB). The latter performed best, and we selected it for our experiments.

VI. Results. Discussion

A. Similarity Measures Evaluation

To test and compare the feature sets defined in section V, we split the dataset into training (80%) and test (20%) sets. For every feature set, the same instance of the classifier (GradientBoostingClassifier) is run. To evaluate the performance of the models, we use weighted average F1 score.

The results for different feature sets (i.e., similarity measures) are presented in Table I.

TABLE I. Evaluation of Different Similarity Measures

Feature set	Precision, %	Recall, %	F1-score, %
shallow	62.42	61.04	60.67
shallow_pos	61.86	60.32	59.92
semantic	62.41	59.28	58.78
distrib	61.22	58.25	57.16
distrib_cosine	60.63	57.69	56.54
distrib_cosine_ext	61.02	58.17	57.22
shallow_semantic	**63.75**	**62.15**	**62.02**
shallow_semantic_distrib	63.72	62.23	62.04
shallow_semantic_distrib_cosine	64.05	62.87	62.68
shallow_semantic_distrib_cosine_ext	**65.73**	**63.90**	**63.66**
shallow_semantic_cosine	63.82	62.55	62.31
shallow_semantic_cosine_ext	64.42	62.95	62.72
all	64.67	63.19	62.98

According to Table I, shallow features perform better than dictionary-based semantic features, which, in their turn, are better than distributional semantic features. It can also be seen that the best scores are achieved on the combination of shallow, dictionary-based and distributional semantic, cosine and extended cosine features. To test whether such combination is significantly better than, for example, the combination of only shallow and dictionary-based semantic features, we conduct a series of pairwise t-tests on the training set during 30-fold cross-validation.

It appears that shallow features perform significantly better (p=0.05) than both dictionary-based semantic and distributional features. Dictionary-based features, in their turn, are significantly better than distributional features (no matter if we add cosine and extended cosine features to the distributional ones or not). POS features do not improve the overall performance at all. Thus, the combination of shallow and dictionary-based semantic features appears to be the best choice for our corpus at the moment.

As it was already mentioned, in this paper we focus on the three types of similarity measures: shallow measures, dictionary-based semantic measures and distributional semantic measures. We do not consider POS-based similarity measures as they have proved to be useless in our experiments. Neither do we consider any combinations of the three types of measures because we intend to compare them. To get a closer look at the performance of the 3 selected types of similarity measures we construct confusion matrices for the corresponding feature sets (see Table II). In Table II precise, loose and non-paraphrases are denoted as 1, 0 and −1 respectively. We adopt such denotation further in this paper.

TABLE II. Evaluation of Different Similarity Measures: Confusion Matrix

		Predicted class		
		1	0	−1
shallow				
True class	1	108	160	13
	0	71	**374**	80
	−1	11	154	**284**
semantic				
True class	1	105	167	9
	0	68	394	63
	−1	8	**196**	245
distrib				
True class	1	**86**	**183**	12
	0	64	399	62
	−1	10	**194**	245

First of all, it is clear from the confusion matrices that all the three types of features tend to mix up neighboring classes of paraphrases (e.g., 1 and 0, 0 and −1), but rarely misclassify 1 as −1 or vice versa. It can also be noted that shallow features are best (among the three considered types of features) at detecting non-paraphrases but worst at detecting loose paraphrases. While semantic dictionary-based (semantic) and distributional semantic (distrib) features tend to mistake non-paraphrases for loose paraphrases, shallow features, on the contrary, more often mistake loose paraphrases for non-paraphrases. Distributional semantic features in general demonstrate the worst performance among the three types of features, especially for precise paraphrases (class 1): they often mistake them for loose paraphrases (class 0).

Thus, the results of the experiments show that simple shallow features perform significantly better than semantic features (which are more complex) on our corpus, and the overall scores are quite low. There are obvious directions for the improvement of the complex semantic features (for example, by introducing more (and richer) semantic resources and using more sophisticated vector space models), but we leave it for further work. At the moment we concentrate on the data to see how the qualities of the text itself affect the performance of different types of similarity measures. Indeed, it is important to take into account the characteristics of the corpus because it can help us tune the existing features and probably come up with new, better ones.

B. Misclassification Analysis and Comparison of the Similarity Measures

To see which pairs of sentences are the most difficult ones for the selected types of similarity measures, we randomly chose 100 misclassified paraphrase pairs for each similarity measure type. These paraphrase pairs were annotated with various linguistic phenomena (see Table III). In table III each value corresponds to the coverage of a particular linguistic phenomenon in a sample of 100 misclassified examples for a particular similarity measure type. The values in the same column do not necessarily sum up to 100% because multiple linguistic phenomena may occur in the same sentence pair.

TABLE III. MISCLASSIFICATION ANALYSIS: LINGUISTIC FEATURES, 100 SAMPLES PER SIMILARITY MEASURE TYPE

Percentage of features in misclassified sentences			
Feature	**distrib**	**semantic**	**shallow**
context knowledge	*16*	*14*	*18*
syntactic synonymy	*26*	*22*	***30***
metonymy	6	5	3
metaphor	2	6	1
phrasal synonymy	11	14	16
different content	**68**	**74**	**64**
synonymy	*22*	*24*	*24*
different time	8	10	14
numeral	6	1	7
reordering	*20*	*14*	*18*

It can be seen from Table III that in the selected samples more than 60% of misclassified sentence pairs contain different information – a phenomenon presumably typical of loose and non-paraphrases. Apparently, such phenomenon represents the main difficulty for all the three types of measures. It allows us to suppose that distinguishing between loose paraphrases and non-paraphrases is probably as difficult as distinguishing between precise and loose paraphrases. Among the three types of measures, for dictionary-based semantic measures the percentage of different content in the misclassified sentences is the highest. Judging from the confusion matrices, it may be supposed that dictionary-based semantic features make so many errors on the sentences with different content due to their general tendency of finding loose paraphrases when there are no paraphrases.

Other less frequent phenomena which take place in the misclassified sentences are context knowledge, synonymy, syntactic synonymy (i.e., the restructuring of some parts of the sentence without altering its meaning) and reordering. It should be noted that syntactic synonymy is harder to tackle for shallow measures than for the other two types of measures, which is in fact expected of such measures.

Table III reflects the qualities of misclassified sentences for each of the similarity measure types separately. To compare them with respect to various linguistic phenomena, we analyzed another distribution: we united 3 samples of 100 annotated sentence pairs and for the obtained sample of 250 sentence pairs for each linguistic phenomenon and for each similarity measure type we calculated the portion of sentence pairs where the corresponding model predicted the wrong paraphrase class (see Table IV).

In Table IV for each linguistic feature absolute frequencies of misclassification and the total number of features in the sample are presented in two rows per feature. The values in Table IV should be interpreted as follows: the last column ("total percentage") displays the percentage of sentence pairs (among the randomly selected 250 sentence pairs) with a particular linguistic feature, and the column entitled "total number" reflects the number of sentence pairs with this features. For example, context knowledge feature occurs in 35 sentence pairs out of 250, i.e., in 14% of sentence pairs. Out of these 35 sentence pairs which require specific knowledge to attribute the pair to a particular paraphrase class, shallow measures fail in 29 sentence pairs, dictionary-based semantic features in 26 sentence pairs and distributional semantic features in 28 sentence pairs. In fact, it means that if the classification of a sentence pair requires context knowledge, then all the three types of similarity measures will misclassify it with high probability. Actually, for any linguistic feature in Table IV all the similarity measures perform unsatisfactorily (> 60% errors).

TABLE IV. MISCLASSIFICATION ANALYSIS: LINGUISTIC FEATURES, PORTIONS OF ERRORS ON THE SAMPLE OF 250 SENTENCE PAIRS (ABSOLUTE NUMBERS AND %)

Number and percentage of misclassified sentence pairs per linguistic feature					
feature	**distrib**	**semantic**	**shallow**	**total number**	**total %**
context knowledge	28	26	29	35	14%
	80%	74.29%	82.86%	100%	
metaphor	6	8	7	8	3,.2%
	75%	100%	87.50%	100%	
numeral	6	8	9	9	3.6%
	66.67%	88.89%	100%	100%	
different content	134	**140**	**126**	176	**70.4%**
	76.14%	79.55%	71.59%	100%	
syntactic synonymy	47	49	50	65	26%
	72.31%	75.38%	76.92%	100%	
metonymy	7	10	9	11	4.4%
	63.64%	90.91%	81.82%	100%	
reordering	29	32	30	46	18.4%
	63.04%	69.57%	65.22%	100%	
synonymy	44	44	40	56	22.4%
	78.57%	78.57%	71.43%	100%	
phrasal synonymy	25	27	27	32	12.8%
	78.13%	84.38%	84.38%	100%	
different time	20	20	21	25	10%
	80%	80%	84%	100%	

The most frequent feature in the sample is that of the different content (70.4%). All the similarity measures misclassify over 70% of sentence pairs with different content, but dictionary-based semantic features do it slightly more often than the others (79.55%), while shallow measures, on the contrary, make fewer mistakes (71.59%). Supposing that different content should only characterize loose paraphrases and non-paraphrases, it means that shallow features are best at distinguishing them, while dictionary-based semantic features are worst (among the three types of similarity measures in question). The second most frequent feature is syntactic synonymy (a phenomenon when the same information is expressed in the sentences using different constituents or the same constituents with different grammatical characteristics). According to Table IV, all the measures perform poorly on the sentence pairs with syntactic synonymy, but shallow measures are slightly worse than the others.

On the whole there is no significant difference in the performance of the similarity measures for the other types of linguistic features. But let us consider the most complicated linguistic phenomena, like metaphor and metonymy. Although they are very infrequent and there is not much related statistics in our sample, it can be seen that distributional semantic measures perform better than the others, while dictionary-based semantic measures are the worst. Apparently, it is due to the fact that distributional measures, unlike dictionary-based ones, cover a wide range of relations. For example, it is a common practice in news texts to refer to a country by the name of its capital (e.g., Russia – Moscow), and distributional measures are likely to capture such associations.

Finally, to get a closer look at the data, we provide some examples of the sentences on which the three types of similarity measures fail to predict the right paraphrase class (5 examples per paraphrase class) in Table V.

TABLE V. MISCLASSIFICATION EXAMPLES

#	Sentences	True class	Predicted class		
			Shallow	Semantic	Distrib + cosine + cosine_ext
1	Осужденному за взрыв на Манежной площади отменили приговор. /The sentence against the convict for the bombing at Moscow's Manezh Square has been cancelled./ Верховный суд отменил приговор за взрыв на Манежной площади. /The Supreme Court has cancelled the sentence for the bombing at Moscow's Manezh Square./	1	0	0	0
2	Порошенко: в Донбассе введут военное положение при атаке на силовиков. /Poroshenko: in Donbass martial law will be imposed to respond to the attack on the security forces./ Порошенко пообещал ответить на наступление ополченцев военным положением. /Poroshenko promised to respond to the attack from the militia with martial law./	1	−1	0	−1
3*	В Альпах разбился снегоход с российскими туристами. /In the Alps, a snowmobile with Russian tourists crashed./ Группа туристов разбилась на снегоходе в Альпах. /A group of tourists crashed on a snowmobile in the Alps./	1	0	0	−1
4	ЦБ РФ потребовал от столичного банка не принимать вклады. /CBR has demanded from the bank of the capital not to take deposits./ ЦБ потребовал у московского банка не принимать вклады. /Central Bank demanded that the Moscow bank does not accept deposits./	1	1	1	0
5	Число погибших при обрушении дома в Мумбаи составило 60 человек. /The death toll in the collapse of the building in Mumbai was 60 people./ В Мумбаи при обрушении дома погибло 60 человек. /In Mumbai 60 people died after the collapse of the building./	1	0	1	0
6	В Москве задержан очередной фигурант “болотного дела”. /In Moscow, another person involved in “swamp case” is arrested./ «Росузник» сообщил о задержании нового фигуранта «болотного дела». /”Rosuznik” informed about the detention of a new person involved in “swamp case”./	0	0	−1	0
7	В аэропорту Казани пассажирский Airbus задел хвостом маяк. /At the airport of Kazan a passenger Airbus hurt a localizer with its tail./ Пассажирский самолет в аэропорту Казани задел хвостом курсовой маяк. /Passenger aircraft at the airport of Kazan hurt a localizer with its tail./	0	1	0	0
8*	В Приднестровье создали Стабфонд для нужд президента и КГБ. /In Transnistria, the Stabilization Fund has been created for the needs of the President and the KGB./ Приднестровье создало Стабфонд за счет российского газа. /Transnistria has created the Stabilization Fund at the cost of the Russian gas./	0	−1	−1	−1
9	В Турции за коррупцию арестованы сыновья трех министров. /In Turkey three sons of ministers have been arrested for corruption./ Турецкая полиция задержала сыновей трех министров. /Turkish police have detained three sons of ministers./	0	−1	0	−1
10	Боевики ИГ взяли на себя ответственность за теракты в Афганистане. /IG militants claimed responsibility for the attacks in Afghanistan./ “ИГ” взяло на себя ответственность за двойной теракт в Афганистане. /“IG” claimed responsibility for the double bombing in Afghanistan./	0	1	1	1
11	После трагедии в Египте запрещены полеты на воздушных шарах. /After the tragedy ballooning is banned in Egypt./ На разбившемся в Египте воздушном шаре не было россиян. /There were no Russians at the crashed balloon in Egypt./	−1	−1	0	0
12	Премьер Киргизии подал в отставку. /Prime Minister of Kyrgyzstan has resigned./ Премьер Киргизии объяснил свою отставку желанием дать дорогу другим. /Prime Minister of Kyrgyzstan explained his resignation with the desire to make way for others./	−1	0	−1	0
13	Рада Украины провалила отставку главы Нацбанка. /Rada failed the resignation of the head of the National Bank./ Верховная Рада провалила назначение нового главы Нацбанка. /Verkhovna Rada failed the appointment of the new head of the National Bank./	−1	0	0	0
14	Минобороны срывает сроки гособоронзаказа. /Defense Ministry breaks deadlines of the defense contracts./ Минобороны отчиталось о размещении гособоронзаказа. /Defense Ministry has presented a report on the defense procurement./	−1	0	0	−1
15	Пожар на территории посольства России в Астане ликвидирован. /Fire on the territory of the Russian Embassy in Astana is eliminated./ В результате пожара в посольстве РФ в Астане никто не пострадал. /At the fire at the Russian Embassy in Astana no one was hurt./	−1	−1	0	−1

Class 1 (Precise Paraphrases). Sentence pairs 1-3 show that all the three measures fail to detect general presupposition (see "convict" and "Supreme Court" in #1: only a convicted person can be a subject to the cancelled sentence, and the court is supposed to cancel the sentence; in #2 "Donbass" as the reference to the place of action in the first sentence is also obvious, especially for the Russian speaker, if the action concerns the martial law imposed to respond to the attack by the militia) as well as syntactic synonymy combined with word and phrase-level synonymy and reordering. We believe that in #3 precise paraphrase class is disputable. For a naïve Russian speaker "tourists" might be identical to "Russian tourists" in a news report due to the presupposition phenomenon, especially if it is a Russian news report, but it is not a general truth. In the #4 example distributional semantic measures apparently do not recognize word-level synonymy. The #5 example includes syntactic synonymy, reordering and words from the same word families (which should be detected by semantic features), and, indeed, only semantic measures predict the true paraphrase class.

Class 0 (Loose Paraphrases). In #8 there is a "difficult" sentence pair: understood metaphorically, the sentences might be considered somewhat similar, however, such understanding requires large amounts of general knowledge and it is extremely hard to teach a machine to distinguish such subtle meanings. Actually #8 is as disputable as #3 as the decision was evidently made by the annotators based on their general presupposition (i.e., prior knowledge about the world). Example #7 confirms the sensitiveness of shallow measures to word overlap: the sentences are of about the same length and are highly overlapping, with the exception of a few words which introduce additional meaning. In #10 the sentences are even more overlapping than in #7, and in this case all the 3 types of measures are mistaken, failing to detect the minor difference between the sentences. In #9 the sentences are smaller and less overlapping. They also contain synonyms and words from the same families. Perhaps, that is why the class 0 is only correctly predicted by dictionary-based model while others misclassified the pair in #9 as non-paraphrase. Pair #6 is similar to #9, but it also contains different named entities which negatively affects the results of the dictionary-based semantic measures (other types of measures are correct here). These named entities express the place of action and the source of information (which is not always important for the reader of a news report) respectively, while the described action is the same in two sentences.

Class −1 (Non-paraphrases). In #14 and #15 it can be seen that dictionary-based semantic measures are also sensitive to word overlap and fail to recognize the difference in the main events described in the sentences. In #12 one of the sentences is much smaller than the other one, and both shallow and distributional measures appear to be sensitive to the fact. #13 is another example of highly overlapping sentences: in the second sentence only one phrase from the first sentence is changed, and it causes the change in the whole meaning because the phrase is antonymous to the original one (see "resignation" and "appointment"). Neither of our types of measures takes semantic contrasts (antonyms and conversives) into account at the moment and, consequently, the predicted classes are incorrect. Example #11 contains some overlapping words, including a named entity, and semantically related words ("трагедии" and "разбившемся"), and, consequently is misclassified by distributional and dictionary-based semantic measures. These two sentences are only recognized as non-paraphrases by the shallow measures.

We can conclude that both paraphrases and non-paraphrases demonstrate quite a high degree of word overlap. It is not surprising that shallow features are the most effective ones. We also believe that high word overlap is naturally characteristic of paraphrases derived from the news texts, which, unlike fiction, are concise and seldom allow for complex variations in the expression of meaning.

Anyway, the overlap in predicate structures which express the main event described in the sentence, and the overlap in modifiers of place, time, etc. are two different phenomena, and they should not be mixed up. Thus, we believe that our semantic models (both dictionary-based and distributional) can be improved by taking into account syntactic structures of the sentences. In the distributional model phrase vectors should also be taken into account, along with the word vectors (as is the case at the moment). It would also be a good idea to introduce weights based on the discriminative power of words and phrases. Dictionary-based model can be improved by considering antonyms as well as synonyms in the sentences.

The analysis of the sentence pairs themselves reveals difficult cases (see examples #3 and #8). The assignment of a particular paraphrase class to such sentence pairs is non-trivial, and naïve Russian speakers are often highly dependent on their prior knowledge when making their decision. Example #8 is especially interesting as it requires an extremely high degree of presupposition to label it as loose paraphrase – if it was annotated by an expert linguist, it is with high probability that the paraphrase class would be the same as the one predicted by the models in question (i.e., class −1). Therefore we plan to introduce another annotation, made by the experts. The level of their agreement is going to serve as an upper bound for the future evaluation of the performance of our models.

VII. Conclusion and Future Work

We have presented the results of the experiments with different types of similarity measures and analyzed their behaviour on the different types of sentences. The experiments were conducted as part of the crowdsourcing project ParaPhraser.ru. In this project we automatically collect Russian sentential paraphrases from the news headlines and work on the development of a paraphrase identification model. The corpus is freely available and quite representational: it can already be used in various studies related to paraphrases and semantics in general (and we use it ourselves to develop a paraphrase identification model). The work on the project is going on, and the corpus is constantly increasing in size.

Our paraphrase corpus includes three classes of sentence pairs: precise paraphrases, loose paraphrases and non-paraphrases. The first two classes are of the main interest to us because they can be used in natural language processing applications like information extraction and text summarization, and we initially supposed that the most difficult task in paraphrase identification would be to distinguish between precise and loose paraphrases. However, the results of our recent experiments disproved this hypothesis: deciding between loose paraphrases and non-paraphrases is also a non-trivial task.

We have compared three classes of similarity measures: shallow measures (based on string/word/phrases overlap), dictionary-based semantic measures (employing external semantic resources) and distributional semantic measures (based on vector space model). The tuning of these measures is left for future work, and in this paper we focused on the analysis of the measures with respect to various linguistic phenomena occurring in sentence pairs derived from the news texts.

Thus, it has been shown that presupposition poses a serious problem for all the considered measures, especially for the shallow ones. We believe that this problem is partly caused by the nature of the annotation (it is crowdsourced at the moment) and that the introduction of "gold" annotation by the experts will help us to solve it.

It has also been shown that in the most difficult sentence pairs with metaphor and metonymy distributional measures are more successful than the others and therefore we should consider the use and the improvement of the corresponding features in our future work. But such difficult cases only occur in the small portion of the data. In most cases, the considered similarity measures misclassify sentences which contain different information (it is also an indication of the "loose paraphrase – non-paraphrase" problem), and here dictionary-based semantic measures are worst, and shallow measures best. Second major source of mistakes is connected with the phenomenon we call "syntactic synonymy", i.e., expressing the same meaning using different constituents, and here, on the contrary, shallow measures are worst, which is not surprising, because they are not intended to capture deep structure of the sentences.

Based on the results of the experiments, we can conclude that all the three considered types of similarity measures are useful for paraphrase identification because their combinations allow us to cover different linguistic phenomena. However, at the moment they all perform poorly on our paraphrase corpus. There is evidently room for the improvement of the measures: for example, both dictionary-based and distributional semantic models can be tuned to recognize synonymy expressed by different syntactic constituents. In distributional measures words can also be weighted according to their discriminative power, etc. As for the paraphrase corpus itself, although experiments with the annotation by naïve speakers can be of certain interest, we still do need experts' annotation because it can reduce the level of presupposition in the sentence pairs and thus make the work for paraphrase identification models easier.

ACKNOWLEDGMENT

The authors acknowledge Saint-Petersburg State University for the research grant 30.38.305.2014.

REFERENCES

[1] P. Achananuparp, X. Hu and Sh. Xiajiong, "The evaluation of sentence similarity measures", *Data Warehousing and Knowledge Discovery*, I.-Y. Song, J. Eder, and T. Nguyen, Eds., vol. 5182 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, pp. 305–316.

[2] 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 the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 238–247.

[3] P. Braslavski, D. Ustalov and M. Mukhin, "A spinning wheel for YARN: user interface for a crowdsourced thesaurus", *in proc. of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 2014, pp. 101–104.

[4] C. Brockett and B. Dolan, "Support vector machines for paraphrase identification and corpus construction", *in proc. of the 3rd International Workshop on Paraphrasing*, 2005, pp. 1–8.

[5] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of semantic distance", *Computational Linguistics*, vol. 32 (1), 2006, pp. 13-47.

[6] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions", *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1 (4), 2007, pp. 300-307.

[7] A. Chitra and S. Kumar, "Paraphrase identification using machine learning techniques", *in proc. of the 12th International Conference on Networking, VLSI and Signal Processing*, 2010, pp. 245–249.

[8] I. Dagan, O. Glickman and B. Magnini, "The PASCAL recognising textual entailment challenge", *in proc. the PASCAL Workshop*, 2005.

[9] D. Das and N. A. Smith, "Paraphrase identification as probabilistic quasi-synchronous recognition", *in proc. of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and of the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 2009.

[10] A. Eyecioglu and B. Keller, "ASOBEK: Twitter paraphrase identification with simple overlap features and SVMs", *in proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 64–69.

[11] S. Fernando, and M. Stevenson, "A semantic similarity approach to paraphrase detection", *in proc. Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloqium*, 2008.

[12] Y. Ji and J. Eisenstein, "Discriminative improvements to distributional sentence similarity", *in proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[13] Z. Kozareva and A. Montoyo, "Paraphrase identification on the basis of supervised machine learning techniques", *in proc. of Advances in Natural Language Processing: 5th International Conference on NLP*, 2006, pp. 524–533.

[14] T. Mikolov, K. Chen, G. S. Corrado and J. Dean, "Efficient estimation of word representations in vector space", Web: http://arxiv.org/abs/1301.3781/, 2013

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

[16] G. Miller and Ch. Fellbaum, "Wordnet: An electronic lexical database", 1998.

[17] E. Pronoza, E. Yagunova and A. Pronoza, "Construction of a Russian paraphrase corpus: unsupervised paraphrase extraction", *in proc. of the 9th Summer School in Information Retrieval and Young Scientist Conference* (in press), 2015.

[18] E. Pronoza and E. Yagunova, "Low-level features for paraphrase identification", *in proc. of the 14th Mexican International Conference on Artificial Intelligence* (in press), 2015.

[19] A. Rajkumar, and A. Chitra, "Paraphrase recognition using neural network classification", *International Journal of Computer Applications* (0975 - 8887), vol. 1 (29), 2010.

[20] V. Rus, Ph. M. McCarthy and M. C. Lintean, "Paraphrase identification with lexico-syntactic graph subsumption", *in proc. of the Twenty-First International FLAIRS Conference*, 2008, pp. 201–206.

[21] H. Schmid, "Improvements in part-of-speech tagging with an application to German", *in proc. of the ACL SIGDAT-Workshop*, Dublin, Ireland, 1995.

[22] G. Sidorov, A. Gelbukh, H. Gómez-Adorno and D. Pinto, "Soft similarity and soft cosine measure: similarity of features in vector space model", *Computación y Sistemas*, vol. 18 (3), 2014, pp. 491–504.

[23] G. Sidorov, H. Gómez-Adorno, I. Markov, D. Pinto and N. Loya, "Computing text similarity using tree edit distance", *NAFIPS'2015* (in press), 2015.

[24] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng and C. D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection", *in proc. of the Conference on Neural Information Processing Systems*, 2011.

[25] S. Wan, M. Dras, R. Dale and C. Paris, "Using dependency-based features to take the "para-farce" out of paraphrase", *in proc. of the Australasian Language Technology Workshop*, 2006, pp. 131–138.

[26] W. Yin and H. Schutze, "Discriminative phrase embedding for paraphrase identification", *in proc. Human Language Technologies: The 2015 Annual Conference of the North Americal Chapter of the ACL*, Denver, Colorado, May 31 – June 5, 2015, pp. 1368-1373.

[27] Y. Zhang and J. Patrick, "Paraphrase identification by text canonicalization", *in proc. of the Australasian Language Technology Workshop*, 2005, pp. 160–166.

[28] A. N. Tikhonov, *Slovoobrazovatelnij Slovar' Russkogo Yazika v Dvuh Tomah: Ok 145000 Slov*. Moscow, Russkiy Yazik, 1985.