View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Helsingin yliopiston digitaalinen arkisto

ParaPhraser: Russian Paraphrase Corpus and Shared Task

Lidia Pivovarova[1], Ekaterina Pronoza[2], Elena Yagunova[2] and Anton Pronoza[3]

[1] University of Helsinki, Helsinki, Finland
[2] St.-Petersburg State University, St.-Petersburg, Russian Federation
[3] Institute for Informatics and Automation of the Russian Academy of Sciences, St.-Petersburg, Russian Federation
lidia.pivovarova@helsinki.fi, katpronoza@gmail.com, iagounova.elena@gmail.com, antpro@list.ru

Abstract. The paper describes the results of the First Russian Paraphrase Detection Shared Task held in St.-Petersburg, Russia, in October 2016. Research in the area of paraphrase extraction, detection and generation has been successfully developing for a long time while there has been only a recent surge of interest towards the problem in the Russian community of computational linguistics. We try to overcome this gap by introducing the project ParaPhraser.ru dedicated to the collection of Russian paraphrase corpus and organizing a Paraphrase Detection Shared Task, which uses the corpus as the training data. The participants of the task applied a wide variety of techniques to the problem of paraphrase detection, from rule-based approaches to deep learning, and results of the task reflect the following tendencies: the best scores are obtained by the strategy of using traditional classifiers combined with fine-grained linguistic features, however, complex neural networks, shallow methods and purely technical methods also demonstrate competitive results.

Keywords: Shared Task, Russian Paraphrase, Paraphrase Detection, Paraphrase Corpus.

1 Introduction

Paraphrase is one of the most problematic concepts in computational linguistics. It has been shown that a narrow definition – "paraphrases must be exactly logically equivalent" – does not cover many cases that are usually considered as paraphrase or quasi-paraphrase (Bhagat and Hovy, 2013). In most practical cases a more relaxed definition of paraphrases is used, e.g. "alternative expressions of the same (or similar) meaning" (Agirre et al., 2015). This notion of similar meaning encompasses a variety of linguistic phenomena, which have a "broad and multi-faceted nature" (Vila et al., 2014). Moreover, in some cases it is hard to distinguish paraphrase and textual entailment, i.e. the implication relation between sentences.

Since it is difficult to work out an exact definition of paraphrase, a data-driven approach might be a reasonable choice. In this case we do not try to give a formal definition of paraphrase but instead lean on native speakers and their judgments whether a particular pair of sentences is a paraphrase or not. In practice, this data-

driven approach requires a construction of large paraphrase corpora with manual or semi-automatic paraphrase annotation, which is obviously a time-consuming task that should be done anew for any given language. On the other hand, recent growth of machine-learning techniques in language processing turns such corpora into valuable resources that can be used to build automatic paraphrase detection systems.

In this paper we present a ParaPhraser project (http://www.paraphraser.ru/) aimed at building of Russian paraphrase corpus, studying of paraphrase phenomena in Russian news and development of automatic paraphrase detection and generation methods (Pronoza and Yagunova, 2015a), (Pronoza andYagunova, 2015b), (Pronoza et al., 2015), (Pronoza et al., 2017). The project was launched in 2014 in St.-Petersburg State University; by the beginning of 2016 we have collected 11 thousand pairs of Russian news titles, which were manually collected as either paraphrase, partial paraphrase or non-paraphrase. The corpus construction process is a combination of automatic paraphrase candidates extraction and manual post-processing of candidate pairs using crowdsourcing. As far as we aware this is the first sentential corpus of Russian paraphrase. From the very beginning the corpus has been publicly available. The current stage of the corpus allowed to perform various research, including linguistic study of paraphrase and study of information flow in news. It also can be used to train automatic paraphrase detection systems, including shared task organized in Fall 2016.

The rest of the paper is organized as follows: in Section 2 we briefly overview related research, including general paraphrase studies, paraphrase corpora and shared tasks; in Section 3 we present the ParaPhraser project and describe the corpus construction process; in Section 4 we present the shared task and its results.

2 Background

2.1 Paraphrase Extraction and Recognition

The detailed survey of paraphrase and textual entailment studies can be found in (Androutsopoulos and Malakasiotis, 2010). We use their exhaustive work as a frame for this section; at the same time, we would like to point out some major changes introduced in the area during the most recent years.

According to (Androutsopoulos and Malakasiotis, 2010), all the tasks related to paraphrases are broken into three main groups: extraction, recognition and generation. Paraphrase extraction is a processing of large corpora aiming at finding paraphrastic sentences or phrases; this is the task we had to solve in the initial step of ParaPhraser corpus generation (see Section 3). Paraphrase recognition means that for a given sentence pair a system should determine whether this is a paraphrase or not; we believe that this task can be solved using ParaPhraser corpus as training data; one of the goals of the shared task, described in Section 4, is to test this assumption. Paraphrase generation, that is a producing of artificial paraphrase for a given sentence, is beyond the scope of this paper, though we are working on this problem as the part of the ParaPhraser project.

Our paraphrase extraction method is based on approach introduced in (Fernando and Stevenson, 2008). They proposed a matrix similarity metric that measures a distance between two sentences based on their word similarity in WordNet. Since a comprehensive Russian WordNet is not currently available we used a synonym dictionary instead of WordNet; we also introduced several modifications into Fernando and Stevenson similarity metric (Pronoza and Yagunova, 2015b).

(Androutsopoulos and Malakasiotis, 2010) listed several methods for paraphrase recognition, including logic-based methods, vector-based methods, those based on surface string similarity, based on syntactic similarity, based on symbolic meaning representation, machine learning methods, and decoding-based methods. Though they mention machine learning as only one method among others, which can be used to combine various features, machine learning methods has become dominating in paraphrase detection area over last years. This does not mean that other methods do not appear in literature; e.g., (Pham et al., 2013) used distributional semantics approach to paraphrase detection. Moreover, it is hard to place a certain approach into single class of the classification. E.g. (Madnani et al., 2012) demonstrated that machine-translation evaluation metrics, such as BLEU, can be effectively used in paraphrase recognition task; most of these metrics utilize surface-string similarity but SVM classifier is used on top of it.

Recent boost in deep learning methods has also affected paraphrase detection studies. Already in 2011, a recursive autoencoder was trained that outperformed state of the art in paraphrase detection task (Socher et al., 2011). An attention-based long short-term memory architecture was used to automatically align pair of sentences and thus measure their similarity (Rocktäschel et al., 2015). A convolutional neural network achieved competitive performance in paraphrase detection task (He et al., 2015).

In the survey conducted by (Androutsopoulos and Malakasiotis, 2010) several natural language processing tasks are mentioned where paraphrase methods can be applied, including question answering, text summarization, information extraction, machine translation, and natural language generation. More recently, even more directions of paraphrase applications have appeared in literature. (Barrón-Cedeño et al., 2013) the authors have demonstrated the importance of paraphrase for plagiarism detection and annotated a plagiarism corpus with paraphrase types. In (Petrović et al., 2012) paraphrase was used for first entity detection task, i.e. to find out the first document that describes a certain news event; they argued that lexical variation is a major obstacle for this task, as well as in number of other tasks, which can be overcome using paraphrase detection techniques. In (Pavlick and Nenkova, 2015) importance of stylistic shifts in paraphrase for genre identification was demonstrated. In (Wieting et al., 2015) the authors used paraphrase corpus to train word embeddings and this improved performance in lexical similarity task. In (Hintz, 2016) it was claimed that paraphrase can be used for stylistic harmonization in multi-document text summarization systems.

Even though the majority of work is done on English data, there is a certain interest in paraphrase research for other languages. For example, in (Eshkol-Taravella and

Grabar, 2014) paraphrastic reformulations in French spoken corpora are studied. In (Nevěřilová, 2014) a paraphrase generation system for Czech was proposed.

There are several publications on paraphrase detection and text reuse for the Russian language, e.g. (Bakhteev et al., 2015), (Khritankov et al., 2015), (Malykh, 2016), however, the amount of research is rather small compared to other languages and to other natural language processing tasks for Russian. One of the missions of the ParaPhraser project is to overcome this gap.

A number of shared tasks on semantic textual similarity have been organized during the last five years as a part of SemEval conferences (Agirre et al., 2012, 2013). The paraphrase detection is very similar to this task, the only difference is that in our task the classification is discrete (paraphrase – non-paraphrase) while in textual similarity the task is to compute a semantic distance using continuous scale. SemEval shared tasks used English and Spanish data (Agirre et al. 2014, 2015). In the most recent shared task there was a sub-task on cross-lingual paraphrase detection (Agirre et al., 2016). There has been organized a special task on semantic similarity in Twitter (Xu et al., 2015). The shared task for paraphrased plagiarism detection has been organized as a part of Russian plagiarism detection shared task (Smirnov et al., 2017) though only one response has been submitted (Zubarev and Sochenkov, 2017). Thus, we can claim that this is a first successful attempt to organize a shared task on Russian paraphrase detection.

2.2 Paraphrase Corpora

There exist a number of available paraphrase corpora. Microsoft Research Paraphrase Corpus (MSRP) (Dolan et al., 2004) is the most known of them. It consists of 5801 pairs of sentences (3900 of them being paraphrases) collected from news clusters. Although it is noted for its loose definition of a paraphrase, its 2-way annotation and high lexical overlap between the sentences (see, for example, Rus et al., 2014, Triantafillou et al., 2016, Liang et al., 2016), it is widely used in paraphrase detection task, and it is the corpus which inspired the development of other paraphrase resources (including our ParaPhraser corpus). MSRP is used as a dataset to monitor state-of-the-art result for paraphrase identification.

Other paraphrase corpora can be classified into several groups depending on the level of paraphrase they cover. Some corpora are purely sentential, while others have additional phrase- or word-level markup. There are also resources which only contain phrasal and word-level paraphrases.

Based on the source of paraphrases, paraphrase corpora can be classified as constructed automatically or manually. The former include parallel multilingual corpora and comparable monolingual corpora, suach as different translations of the same texts, news texts, texts on similar topics, e.g., from the social networks or students' answers to the questions, social media, Wikipedia, different descriptions of the same videos.

Sentential Corpora. One of the oldest sentential corpora known to us is the KMC corpus (Knight and Marcu, 2002) collected from pairs of texts and their summaries.

User Language Paraphrase Corpus (McCarthy and McNamara, 2008) is collected from student paraphrases of biology textbook sentences. Question Paraphrase Corpus (Bernhard and Gurevych, 2008) includes sentences pairs derived from WikiAnswers and annotated by social media users. Microsoft Research Video Description Corpus (Chen and Dolan, 2011) is collected from short descriptions of videos annotated on the Amazon Mechanical Turk crowdsourcing platform.

Regneri and Wang corpus (Regneri et al., 2014) is collected from summaries of TV show episodes. Twitter Paraphrase Corpus (Xu et al., 2013) is derived from tweets corresponding to the same events (referring to the same date and mentioning the same named entity). Student Response Analysis Corpus (Dzikovska et al., 2013), is collected from students' answers to explanation and definition question. Semantic Textual Similarity Corpus (Agirre et al., 2013) is collected from several sources including news texts, Framenet-WordNet glosses and OntoNotes-WordNet glosses.

Non-English sentential paraphrase corpora known to us are Japanese Paraphrase Corpus for Speech Translation (Shimohata et al., 2004), consisting of sentences derived from travel conversation and versions of them paraphrased by humans, and Turkish Paraphrase Corpus (Demir et al., 2012), covering both sentence- and word- and phrase-level paraphrases, and derived from several sources: translations of a famous novel, subtitles, translations from an English-Turkish parallel corpus, and articles from a news website. More recently, another Turkish paraphrase corpus has been compelled by (Eyecioglu and Keller, 2016).

Phrasal Corpora. The corpus compiled by (Cohn et al., 2008) is derived from three different sources: the multi-translation Chinese corpus (mtc), Jules Verne's "20,000 leagues under the sea" novels and MSRP (with non-paraphases).

WiCoPaCo (Max and Wisnewski, 2010) is a corpus of French paraphrases collected from Wikipedia's revision history. WRPA (Vila et al., 2010) is another corpus based on Wikipedia and taking advantage of its structure. Unlike WiCoPaCo it captures only paraphrases of specific relationions (authorship, person-date of birth relation, etc.). The SEMILAR Corpus (The SEMantic SimILARity Corpus, (Rus et al., 2012)) is based solely on MSRP, enriched with word level similarity and alignments.

The Paraphrase Database developed by (Ganitkevitch and Callison-Burch, 2014) is a rich paraphrase resource, which includes billions of paraphrase pairs. It is collected for more than 20 languages, including Russian, from bilingual parallel corpora. The authors use a language independent method to extract paraphrases from parallel bilingual texts: paraphrases are found in a single language by "pivoting" over a shared translation in another language. This approach was introduced by (Bannard and Callison-Burch, 2005) and has been successfully applied by many researchers. PPDB includes lexical, phrasal and syntactic paraphrases, all of which are annotated with metrics from machine translation.

3 The ParaPhraser Project

There have been no publicly available paraphrase resources for the Russian language known to us, with the only exception of the dataset published by (Ganitkevitch and Callison-Burch, 2014) as part of The Paraphrase Database project. The latter includes paraphrases on the word-, phrase- and syntactic levels, but it lacks information on the context of paraphrases. That is why we have constructed a sentential paraphrase corpus as part of our ParaPhraser project. The project is aimed at studying paraphrase phenomenon in Russian, including paraphrase extraction, paraphrase corpora construction and building paraphrase identification and generation models. Our results of solving paraphrase generation problem are available in the form of RESTful API service (https://paraphraser.ru/api/form), and the collected paraphrase corpus is also freely available on our website (http://paraphraser.ru/download). The corpus is not intended to be a general-purpose one. It consists of sentential paraphrases, extracted from news headlines, since news analysis is our primary interest, with the focus on such practical tasks as information extraction and text summarization.

To build the corpus we use a two-step procedure: first, automatic collection of candidate pairs and then manual annotation using crowdsourcing. We now describe both stages in more details.

3.1 The Construction Process

In the ParaPhraser project, we collect sentence pairs in real time. We parse web sites of several Russian news agencies and extract headlines of the articles. The headlines, as in the strategy proposed by (Wubben et al., 2009), are compared to each other, and paraphrase candidates are extracted using a similarity metric which extends the unsupervised matrix similarity metric proposed by (Fernando and Stevenson, 2008) and is also a variant of soft cosine measure (Sidorov et al., 2014). A detailed description of metric calculation can be found in (Pronoza and Yagunova, 2015b). We include in the corpus pairs of sentences with the similarity metric value larger than a certain threshold. To provide more negative instances, we also include in the corpus a small random portion of sentence pairs with similarity metric value below the threshold.

3.2 Crowdsourcing

Potential paraphrases are annotated via our online interface[1]. The annotators are native speakers of Russian. Most of them are naïve speakers but there are also expert linguists and students of linguistics. Two sentences at a time are shown to an annotator and she/he decides whether the sentences convey the same meaning (1), similar meanings (0) or different meanings (-1). There are no specific instructions; instead, we let them use their own judgment and intuition. We introduce an

1 http://paraphraser.ru/scorer

entertainment element into the tedious annotation process: the annotators are shown funny pictures and/or facts at random intervals and are encouraged to work further. Inter-annotator agreement, calculated as Kohen's Cappa for all pairs of annotators, does not exceed 0.6.

When calculating resulting paraphrase classes, we only consider sentence pairs annotated by at least 3 users. We discard sentence pairs with opposite judgments (-1 and 1). Paraphrase class of a pair of sentences in the corpus is calculated as the median of all the scores given to this pair by the annotators (in case of ties the values are round down to the previous integers (0.5 to 0 and -0.5 to -1).

3.3 Evaluation

Due to our paraphrase construction method only small subset of instances classified as negative by the algorithm is selected for manual assignment, which is not sufficient to compute recall. Thus we use precision to evaluate the quality of the unsupervised similarity metric for corpus construction. Precision of the metric on the current corpus, i.e. the training dataset used for the Shared Task, is 79.92%. Previously we evaluated our metric used for corpus construction (Pronoza et al., 2015c) when the corpus consisted of about 5 thousand sentence pairs, and metric precision was 80.24%. Such results are quite promising compared with the original metric by Fernando and Stevenson that achieved 75.2% against MSRP.

4 Shared Task

4.1 The Task

The task input was a set of sentence pairs collected from news headlines and manually annotated by three native speakers as precise paraphrase, near paraphrase and non-paraphrase, as it is described in the previous section. We used approximately 7 and 2 thousand pairs for the training and test sets respectively. Both training and test sets are freely available[2].

The ParaPhraser corpus has been freely available from the very beginning, which means that all manually annotated data immediately became public. Only when we decided to organize the shared task we stopped publishing data to collect a test set. Thus, the training and the test sets are collected during different time periods and potentially annotated by different people. Some participants of the shared task noticed that cross-validation results were slightly better than results obtained on the test set, which can be explained by the fact that the test set was not a random sample from the data.

The shared task consisted of two sub-tasks:

Task 1. Three-class classification: given a pair of sentences, to predict whether they are precise paraphrases, near paraphrases or non-paraphrases.

2http://www.paraphraser.ru/download/

Task 2. Binary classification: given a pair of sentences, to predict whether they are paraphrases (whether precise or near paraphrases) or non-paraphrases.

For each task we allowed standard and non-standard runs. In standard runs participants could not use any corpora but ParaPhraser or any derivatives from these external corpora (such as embeddings). However, we allowed to use any language processing tools or manually compiled dictionaries in standard runs. Any resources were allowed for non-standard runs.

Submissions of the participants were evaluated using accuracy and F1-score (F1-micro and F1-macro for three-class classification task).

4.2 Baselines

We provided two baselines for both tasks (2-class and 3-class classification). The first baseline assigns *random* class to each pair of sentences. The second baseline (*baseline2*) is a bit more complicated and consists of two steps. First, we conduct stemming of all words consisting of more than two characters by cutting off two characters from the end of a word. Then we compute a number of the overlapping words. For two-way classification a pair is classified as a paraphrase if more than a half of words from the longer sentence are mentioned in the shorter one. For 3-way classification we consider that the pair is a near-paraphrase if overlap of words is between 33% and 50% and precise paraphrase pair in case the overlap is more than 50%. Despite the simplicity of the technique the results appeared not to be the worst.

4.3 Results

For each task each participant might submit 20 standard and 20 non-standard runs. Since none of the participants made that many submissions we can assume that all participants submitted as many different responses as they wanted.

In total, 16 teams registered to the shared task, 11 submitted at least one result. The organizers submitted baseline results and an additional algorithm, described in the next section. The final results are presented in Tables 1-4[3]. For each team we present only the best result.

As can be seen from the tables, three-way classification is a more difficult task than two-way classification, for those systems that participated in both tasks the difference is up to 15% in accuracy and up to 30% in F-measure. The difference between standard and non-standard runs is not that high, which might be explained by the nature (and rather small amount) of our data: the sentences in the ParaPhraser corpus are highly overlapping, which is common for corpora constructed from news texts, and simple shallow methods are usually more successful when tried against such data.

The participants of the task used a wide variety of techniques, from rule-based approaches to deep learning, and results of the task reflect the following tendencies: the best scores are obtained by the strategy of using traditional classifiers combined with fine-grained linguistic features, however, complex neural networks, shallow

3 In some cases we don't know, what method was used.

methods and purely machine learning methods also demonstrate competitive results[4]. The best results for two-way classification are slightly lower than English state of the art: the best result reported on ACL Anthology wiki page[5] yields accuracy 80.4% and F1-measure 85.9% though it is hard to compare results obtained on different corpora.

The papers, submitted to this volume present a variety of methods:

- Rule-based semantic parser (Boyarsky and Kanevsky, 2017)
- SVM or Random Forest classifiers on top of thesaurus-based similarity features (Loukachevitch et al. 2017)
- SVM classifier on top of word and character unigrams, bigrams and trigrams (Eyecioglu and Keller, 2017)
- Gradient Boosting classifier on top of features obtained from existing toolkits, including machine translation and similarity detection tools for the English language (Kravchenko, 2017)
- Convolutional neural networks (Maraev et al. 2017)

Table 1. Results: 3-way classification, standard run

Team	Accuracy	F1 (macro)	Method
Team3448	0.5901	0.5692	Classifier + linguistic features
AsoBek	0.5732	0.5557	Classifier + surface features
Penguins	0.5721	0.4443	Textula similarity based on word embeddings
MLforNLP	0.5695	0.5437	Technological approach (including translation into English)
True Positive	0.5631	0.5382	Classifier + linguistic features
dups	0.5478	0.5175	Neural networks
Baseline2	0.5325	0.5096	
DHL	0.4881	0.4483	Neural networks?
PhraseAnalog	0.4522	0.4344	Rule-based system
Team	0.4068	0.3699	
Random	0.3439	0.3341	

4These are observations done during the shared task workshop at the AINL 2016 conference. Unfortunately, not all participants submitted a paper though some presentations are available on the conference webpage: http://ainlconf.ru/2016/materials

5 https://aclweb.org/aclwiki/index.php?title=Paraphrase_Identification_(State_of_the_art)

Table 2. Results: 2-way classification, standard run

Team	Accuracy	F1	Method
dups	0.7459	0.8044	Neural networks
Team3448	0.7448	0.8078	Classifier + linguistic features
NLX	0.7274	0.7880	Neural networks
AsoBek	0.7211	0.7873	Classifier + surface features
True Positive	0.7179	0.7656	Classifier + linguistic features
MLforNLP	0.7153	0.7853	Technological approach (including translation into English)
DHL	0.6292	0.7325	Neural networks
Baseline2	0.5858	0.5094	
Random	0.4966	0.5403	
Penguins	0.4702	0.2170	Textual similarity based on word embeddings

Table 3. Results: 3-way classification, non-standard run

Team	Accuracy	F1 (macro)	Method
True Positive	0.6181	0.5838	Classifier + linguistic features
dups	0.5969	0.5680	Neural networks
Team3448	0.5853	0.5642	Classifier + linguistic features
L533	0.5832	0.5567	
DHL	0.4099	0.3576	Neural networks?

Table 4. Results: 2-way classification, non-standard run

Team	Accuracy	F1	Method
True Positive	0.7739	0.8110	Classifier + linguistic features
dups	0.7665	0.7982	Neural networks
Team3448	0.7343	0.7827	Classifier + linguistic features
L533	0.6926	0.7794	
DHL	0.5605	0.6916	Neural networks

4.4 Experiments

The task organizers submitted runs for both tasks as Team3448. Our approach towards the problem of paraphrase detection is based on the use of three types of sentence similarity measures as features in the paraphrase classification task (Pronoza and Yagunova, 2015a): 1) surface, or shallow, similarity measures based on the overlap of n-grams, words and characters in the sentences; 2) semantic similarity measures that cover synonymy relations and derivation morphology; 3) distributional measures that use vector representations of words and phrases.

In total, we use 24 shallow features, 11 semantic features and 45 distributional features. Most of the features are described in (Pronoza and Yagunova, 2015a), the others are distributional features with *phrase* embeddings with discriminative weights and 3-nearest neighbours smoothing for unknown words.

We submitted both standard and non-standard runs. Our models for the non-standard runs were built using all the described features. In the standard setting we cut off distributional features since they used external corpora. We tried two classifiers: SVM and Gradient Tree Boosting. Parameters of SVM and GTB were optimized on the development set (20% of the training set).

This approach achieved quite competitive results and even obtained the 1st place in the standard run of Task 1. Surprisingly, results of our standard runs are better than those of non-standard runs (the former use external resources and richer feature sets than the latter ones). This is similar to a general tendency, presented in Tables 1–4, where non-standard runs gain only little improvement.

5 Conclusion

We presented a freely available ParaPhraser corpus and the first shared task on Russian Paraphrase detection. We demonstrate that the corpus can be used for such task, which means that it is potentially useful for practical applications that require paraphrase identification step, such as cross-document text summarization or information extraction. The shared task results demonstrate that paraphrase detection methods developed for other languages may be applied to Russian and yield results only little worse than the English state of the art.

According to the results of the Shared task, various methods, from rule-based systems to deep learning, can be used for paraphrase detection, and most of them are quite successful at the task in question. As our dataset is not large, we expected a traditional (classifier + fine-grained features) approach to achieve best scores, and the results of the task met our expectations. However, deep learning approach also obtained high results (and the first place in one of the subtasks), and other methods, both surface and complex ones, appeared to be competitive.

We continue collecting data for the corpus. In total, we have already collected about 11 thousand pairs of sentences, which is 2 thousand more than we had during the shared task evaluation campaign. These 2 thousand are not yet publicly available since we plan to use part of it as a test set in the next shared task.

Though the shared task was quite successful there are also lessons learned, that we will use in the next shared tasks. First, we should have asked all participants to submit a short description of their method, so that we knew which approaches were tried even if the team decided not to submit the paper. Second, we should try to balance training and test set, so that training set contains some sentence pairs annotated by the same annotators as the test set and during the same period of time.

Another idea is to use much larger test set, where some pairs would be manually annotated and used to compute the evaluation measures and some pairs would be only automatically collected. These would serve for two goals: make it more difficult to optimize systems to a particular test set and reduce human efforts in annotation since the pairs on which all participating systems agree might be added to the corpus without human annotation.

6 References

1. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Wiebe, J. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. Proceedings of SemEval 2014 (2014).
2. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G.; Uria, L., Wiebe, J. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. Proceedings of SemEval 2015 (2015).
3. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R.; Rigau, G., Wiebe, J. Semeval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation Proceedings of SemEval 2016 (2016).
4. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. Proceedings of SemEval 2012 (2012).
5. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo. W. *SEM 2013 shared task: Semantic Textual Similarity, Proceedings of *SEM 2013 (2013).
6. Androutsopoulos, I., Prodromos Malakasiotis, P. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research, v. 38: 135–187 (2010).
7. Bakhteev, O., Kuznetsova, R., Romanov, A., Khritankov, A. A monolingual approach to detection of text reuse in Russian-English collection. In Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), pp. 3–10, IEEE (2015).
8. Bannard, C., Callison-Burch, Ch.: Paraphrasing with Bilingual Parallel Corpora. In Proceedings of the 43rd Annual Meeting of the ACL, pp. 597–604 (2005).
9. Barrón-Cedeño, A., Vila, M., Martí, M. A., Rosso, P. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics, 39(4), 917–947 (2013).
10. Bernhard, D., Gurevych, I. Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. In Proceedings of the ACL'08 3rd Workshop on Innovative Use of NLP for Building Educational Applications, p. 44–52 (2008).
11. Bhagat, R., Hovy, E. What is a paraphrase? Computational Linguistics, 39 (3): 463–472 (2013).
12. Boyarsky K., Kanevsky E. Effect of Semantic Parsing Depth on the Identification of Paraphrases in Russian Texts. CCIS 789 (this volume)

13. Chen, D. L., Dolan, W. B. Collecting Highly Parallel Data for Paraphrase Evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 190–200, Portland, Oregon, USA (2011).
14. Cohn, T., Callison-Burch, C., Lapata, M. Constructing corpora for the development and evaluation of paraphrase systems. Computational Linguistics, 34(4), 597–614 (2008).
15. Demir, S., El-Kahlout, ˙Il. D., Unal, E., Kaya, H. Turkish Paraphrase Corpus. In proceedings of LREC'2012, pp. 4081–4091, (2012).
16. Dolan, B., Quirk, C., Brockett, C. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of the 20th international conference on Computational Linguistics (p. 350). Association for Computational Linguistics (2004).
17. Dzikovska, M. O., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I. and Dang, H.T. SemEval - 2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA (2013).
18. Eshkol-Taravella, I., Grabar, N. Paraphrastic Reformulations in Spoken Corpora. In International Conference on Natural Language Processing, pp. 425–437, Springer International Publishing (2014).
19. Eyecioglu, A., Keller, B. Constructing A Turkish Corpus for Paraphrase Identification and Semantic Similarity. In Proceedings of the CICLIng, 17th International Conference on Intelligent Text Processing and Computational Linguistics. Lecture Notes in Computer Science, vol. 9623, pp. 562–574 (2016).
20. Eyecioglu, A., Keller, B. Knowledge-lean Paraphrase Identification Using Character-Based Features. CCIS 789 (this volume)
21. Fernando, S., Stevenson, M. A semantic similarity approach to paraphrase detection. In Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics, pp. 45–52, (2008).
22. Ganitkevitch, J., Callison-Burch, Ch.: The Multilingual Paraphrase Database. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA). Reykjavik, Iceland (2014).
23. Ganitkevitch, J., Van Durme, B., Callison-Burch, Ch. PPDB: The Paraphrase Database. In HLT-NAACL, pp. 758–764 (2013).
24. He, H., Gimpel, K., Lin, J. Multi-perspective sentence similarity modeling with convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1576–1586 (2015).
25. Hintz, G. Data-driven Paraphrasing and Stylistic Harmonization. In Proceedings of NAACL-HLT, pp. 37–44 (2016).
26. Khritankov, A., Botov, P., Surovenko, N., Tsarkov, S., Viuchnov, D., Chekhovich, Y. Discovering Text Reuse in Large Collections of Documents: a Study of Theses in History Sciences. In Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), pp. 26–32, IEEE (2015).
27. Knight, K., Marcu, D. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. In: Artificial Intelligence, vol. 139 (1): 91–107 (2002).
28. Kravchenko D. Paraphrase Detection Using Machine Translation and Textual Similarity Algorithms. CCIS 789 (this volume)

29. Liang, Ch., Paritosh, P., Rajendran, V., Forbus, K. D. Paraphrase Identification with Structural Alignment. Proceedings of the 16th International Joint Conference on Artificial Intelligence, pp. 2859–2865 (2016).
30. Loukachevitch N., Shevelev A., Mozharova V., Dobrov B. and Pavlov A. RuThes Thesaurus in Detecting Russian Paraphrases. CCIS 789 (this volume)
31. Madnani, N., Tetreault, J., Chodorow, M. Re-examining machine translation metrics for paraphrase identification. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 182–190, Association for Computational Linguistics (2012).
32. Malykh, V. Robust Word Vectors for Russian Language. In Proceedings of Artificial Intelligence and Natural Language AINL FRUCT 2016 Conference, Saint-Petersburg, Russia, 10–12 November 2016, pp.95–98 (2016).
33. Maraev V., Saedi Ch., Rodrigues J., Branco A., and Silva J. Character-level Convolutional Neural Network for Paraphrase Detection and other Experiments. CCIS 789 (this volume)
34. Max, A., Wisniewski, G. Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History, LREC 2010, Valetta, Malta (2010).
35. McCarthy, Ph. M., McNamara, D. S. The User-Language Paraphrase Corpus. Cross-Disciplinary Advances in Applied Natural Language Processing (2008).
36. Nevěřilová, Z. Paraphrase and textual entailment generation in Czech. Computación y Sistemas, 18 (3): 555–568 (2014).
37. Pavlick, E., Nenkova, A. Inducing lexical style properties for paraphrase and genre differentiation. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 218–224 (2015).
38. Petrović, S., Osborne, M., Lavrenko, V. Using paraphrases for improving first story detection in news and Twitter. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 338–346. Association for Computational Linguistics (2012).
39. Pham, N., Bernardi, R., Zhang, Y. Z., Baroni, M. Sentence paraphrase detection: When determiners and word order make the difference. In Proceedings of the Towards a Formal Distributional Semantics Workshop at IWCS 2013, pp. 21–29 (2013).
40. Pronoza E., Yagunova E., Kochetkova N. Sentence Paraphrase Graphs: Classification Based on Predictive Models or Annotators' Decisions?. In: Sidorov G., Herrera-Alcántara O. (eds) Advances in Computational Intelligence. MICAI 2016. Lecture Notes in Computer Science, vol 10061. Springer, Cham (2017).
41. Pronoza, E., Yagunova, E. Comparison of sentence similarity measures for Russian paraphrase identification. In Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), pp. 74–82, IEEE (2015a).
42. Pronoza, E., Yagunova, E. Low-Level Features for Paraphrase Identification. In: Sidorov, G., Galicia-Haro, Sofía N. (eds.) MICAI 2015. LNCS, vol. 9413, pp. 59–71. Springer, Cham (2015b).
43. Pronoza, E., Yagunova, E., Pronoza, A. Construction of a Russian Paraphrase Corpus: Unsupervised Paraphrase Extraction. Proceedings of the 9th Russian Summer School in Information Retrieval, August 24–28, 2015, Saint-Petersburg, Russia, (RuSSIR 2015, Young Scientist Conference), Springer CCIS (2015).
44. Regneri, M., Wangy, R., Pinkal, M. Aligning Predicate-Argument Structures for Paraphrase Fragment Extraction. LREC 2014: 4300–4307 (2014).

45. Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., Blunsom, P. Reasoning about entailment with neural attention. arXiv preprint arXiv:1509.06664. (2015).
46. Rus, V., Banjade, R., Lintean, M. On Paraphrase Identification Corpora. LREC'2014, pp. 2422-2429 (2016).
47. Rus, V., Lintean, M., Moldovan, C., Baggett, W., Niraula, N., Morgan, B. The SEMILAR Corpus: A Resource to Foster the Qualitative Understanding of Semantic Similarity of Texts, In Semantic Relations II: Enhancing Resources and Applications, The 8th Language Resources and Evaluation Conference (LREC 2012), May 23-25, Instanbul, Turkey (2012).
48. Shimohata, M., Sumita, E., Matsumoto, Y. Building a Paraphrase Corpus for Speech Translation. In Proceedings of 4th international conference on language resources and evaluation (LREC) (2004).
49. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D. Soft similarityand soft cosine measure: similarity of features in vector space model. Computación y Sistemas, vol. 18 (3): 491–504 (2014).
50. Smirnov, I., Kuznetsova, R., Kopotev, M., Khazov, A., Lyashevskaya, O., Ivanova, L., Kutuzov, A. Evaluation Tracks on Plagiarism Detection Algorithms for the Russian Language. Dialog 2017 (2017).
51. Socher, R., Huang, E. H., Pennin, J., Manning, C. D., Ng, A. Y. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In Advances in Neural Information Processing Systems, pp. 801–809 (2011).
52. Triantafillou, E., Kiros, J. R., Urtasun, R., Zeme, R. Towards Generalizable Sentence Embeddings. In Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 239–248, Berlin, Germany (2016).
53. Vila, M., Martí, M. A., Rodríguez, H. Is this a paraphrase? What kind? Paraphrase boundaries and typology. Open Journal of Modern Linguistics, 4 (01), 205 (2014).
54. Vila, M., Rodriguez, H., Marti, M. A. WRPA: A System for Relational Paraphrase Acquisition from Wikipedia. Procesamiento del Lenguaje Natural, Revista No 45, septiembre 2010, pp. 11–19 (2010).
55. Wieting, J., Bansal, M., Gimpel, K., Livescu, K. Transactions of the Association for Computational Linguistics, vol. 3, pp. 345–358 (2015).
56. Wubben, S., van den Bosch, A., Krahmer, E., Marsi, E.: Clustering and Matching Headlines for Automatic Paraphrase Acquisition. In: Proceedings of the 12th European Workshop on Natural, language Generation, pp. 122–125, Athens, Greece (2009).
57. Xu, W., Callison-Burch, Ch., Dolan, W. B. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). Proceedings of SemEval (2015).
58. Xu, W., Ritter, A., Grishman, R. Gathering and Generating Paraphrases from Twitter with Application to Normalization. Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, August 2013, Sofia, Bulgaria, pp. 121–128 (2013).
59. Zubarev, D. V., Sochenkov, I. V. Paraphrased Plagiarism Detection Using Sentence Similarity. Dialog 2017 (2017).