# Shared Task on Sentence Paraphrase Detection for the Russian Language

Lidia Pivovarova[1], Ekaterina Pronoza[2], Elena Yagunova[2]

[1]University of Helsinki; [2]St.-Petersburg State University

{lidia.pivovarova, katpronoza, iagounova.elena}@gmail.com

**Problem overview**

Sentence-level paraphrase detection is identification of sentences that have similar meaning but not necessary similar form. This procedure may serve as a crucial initial step for various text processing tasks, such as text summarization, document clustering and plagiarism detection. From technical point of view this is a complex task that involves text analysis on various levels, from morphology to semantics.

Paraphrase detection has recently caused a number of publications, e.g. [1],[2],[3],[4],[5],[6],[7] to name a few; the growing interest to the topic may be partially explained by recent achievements in distributional semantic methods that are able to compute word and text similarity without manually compelled semantic dictionaries. The number of shared tasks on semantic text similarity has been organized as a part of SemEval conferences ([8],[9][10],[11], see also the task Wiki page: http://ixa2.si.ehu.es/stswiki/index.php/). In these tasks, the participating systems take as an input a pair of sentences and produce as an output a similarity score for the pair; most of evaluations so far there done on English datasets.

The Russian language, as a language with a rich morphology and a free word order, may provide an interesting data for paraphrase detection techniques. There are several publications on paraphrase detection and text reuse for the Russian language, e.g. [12],[13],[14] (see also a series of papers by the task organizers, e.g. [15],[16]). The shared task on word-level semantic similarity [17] was organized as a part of Dialogue Evaluation workshop and attracted six participating teams. Nevertheless, the amount of work invested into sentence-level paraphrase detection is less than for the other text analysis task for Russian, most probably because this task required a substantial amount of testing and training data, which have not existed for Russian.

This shared task is an initiative aiming to boost research on the paraphrase detection for Russian. A core dataset for the task is ParaPhraser [18], a freely available corpus of Russian sentence pairs manually annotated as paraphrases, near-paraphrases and non-paraphrases. Each candidate pair was collected from news headlines and then manually annotated by three native speakers. The current size of the corpus is 7000 pairs, these data will be used as a training set. The test set is being currently collected using the same crowdsourcing procedure and will become a part of the general corpus after the end of the shared task.

**Tasks description**

The shared task will follow the standard procedure: the participating teams will take as an input a pair of sentences and return as a response the similarity class.

We distinguish three possible classes: precise paraphrase, near paraphrase and non-paraphrase. Examples of sentences of these three classes are shown in Table 1.

| Paraphrase class | Example |
|---|---|
| Precise paraphrase | КНДР аннулировала договор о ненападении с Южной Кореей. |
| | *DPRK annulled the non-aggression treaty with South Korea.* |
| | КНДР вышла из соглашений о ненападении с Южной Кореей. |
| | *DPRK withdrew from the non-aggression agreement with South Korea.* |
| Near paraphrase | ВТБ может продать долю в Tele2 в ближайшие недели. |
| | *VTB might sell its shares in TELE2 in the nearest weeks.* |
| | ВТБ анонсировал продажу Tele2. |
| | *VTB announced the sale of TELE2.* |
| Non-paraphrase | В главном здании МГУ загорелась столовая. |
| | *The student canteen has lit in the main building of MSU.* |
| | Из главного здания МГУ эвакуированы около 300 человек. |
| | *About 300 people are evacuated from the main building of MSU.* |

Table 1. Examples of paraphrase classes.

The shared task will consist of two tasks.

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

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

Participants may submit “standard” runs, which use as training data only the ParaPhraser corpus, and “non-standard” runs, which may use any other data. “Standard” and “non-standard” run will be evaluated separately.

**Training data**

Training data are available at http://www.paraphraser.ru/download/get?file_id=1

It contains about 7 thousand sentence pairs labeled as precise paraphrases, near paraphrases and non-paraphrases.

**Testing data**

Testing data will be made available in the same format as the training data, with paraphrase classes omitted. In the submitted files, paraphrase classes (0, 1 or -1 – for task 1; 0 or 1 – for task 2) should be inserted as values into the corresponding tags.

Detailed formats of the testing data and participant responses are given in Appendices A and B respectively.

**Evaluation**

The quality of submitted results for task 1 for both “standard” and “non-standard” runs will be assessed using Accuracy and two variants of average F1-score: with micro- and macro-averaging.

The quality of submitted results for task 2 for both “standard” and “non-standard” runs will be assessed using Accuracy and F1-score.

Accuracy is proportion of correctly classified sentence pairs in all pairs in the test set:

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + t_n + f_n},$$

where

- $t_p$ is the number of true positives (i.e., the number of correctly classified sentence pairs),
- $t_n$ is the number of true negatives,
- $f_p$ is the number of false positives and
- $f_n$ is the number of false negatives.

F1-score is the harmonic mean of Precision ( $P$ ) and Recall ( $R$ ):

$$F_1 = \frac{2 \cdot P \cdot R}{P + R},$$

$$P = \frac{t_p}{t_p + f_p},$$

$$R = \frac{t_p}{t_p + f_n}.$$

Precision, Recall and F1-score can only be calculated for a single class, i.e. these measures are suitable for binary classification problem. In a multi-class classification task like **task 1** average measures are used. In **task 1** results are to be evaluated using macro and micro F1-score.

Macro-averaged F1-score is calculated as the average of F1-scores of all the classes:

$$F_{macro} = \frac{1}{n} \sum_{i=1}^{n} F_1^i,$$

where $F_1^i$ denotes F1-score for the $i$ th class, and $n$ is the total number of classes (or binary classification problems).

Micro-averaged F1-score is calculated as the harmonic mean of micro-averaged Precision and Recall. The latter are compute using the sums of true positives, false positives and false negatives:

$$F_{micro} = \frac{2 \cdot P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}},$$

$$P_{micro} = \frac{\sum_{i=1}^{n} t_{p_i}}{\sum_{i=1}^{n} (t_{p_i} + f_{p_i})},$$

$$R_{micro} = \frac{\sum_{i=1}^{n} t_{p_i}}{\sum_{i=1}^{n} (t_{p_i} + f_{n_i})},$$

where $t_{p_i}$, $f_{p_i}$ and $f_{n_i}$ denote the number of true positives, false positives and false negatives for the $i$ th class, and $n$ is the total number of classes (or binary classification problems).

**Important Dates**

The tentative timeline for the shared tasks is as following:

- June 2016: the first call for participation; the training data are already available on the corpus webpage (http://www.paraphraser.ru/download/)
- 1st September 2016: the second call for participation
- 1st October 2016: test set available
- 10th October 2016: system responses deadline
- 12th October 2016: official results announces
- 11-12 November: the workshop with the participants' presentations and the final discussion (as a part of AINL 2016 program, http://ainlconf.ru/)
- End of December 2016: the full paper deadline

**Publication**

We invite participants of the shared task to submit papers describing their approach to the problem of paraphrase detection. Accepted papers will be published in a volume indexed by international databases (details are to be announced soon).

**References**

1. Barrón-Cedeño, A., Vila, M., Martí, M. A., & Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics, 39(4), 917-947.

2. Cohn, T., Callison-Burch, C., & Lapata, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. Computational Linguistics, 34(4), 597-614

3. Dolan, B., Quirk, C., & Brockett, C. (2004). 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.

4. Fernando, S., & Stevenson, M. (2008). 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).

5. Madnani, N., Tetreault, J., & Chodorow, M. (2012, June). 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.

6. Pham, N., Bernardi, R., Zhang, Y. Z., & Baroni, M. (2013). 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).

7. Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In Advances in Neural Information Processing Systems (pp. 801-809).

8. Eneko Agirre; Carmen Banea; Claire Cardie; Daniel Cer; Mona Diab; Aitor Gonzalez-Agirre; Weiwei Guo; Inigo Lopez-Gazpio; Montse Maritxalar; Rada Mihalcea; German Rigau; Larraitz Uria; Janyce Wiebe. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. Proceedings of SemEval 2015

9. Eneko Agirre; Carmen Banea; Claire Cardie; Daniel Cer; Mona Diab; Aitor Gonzalez-Agirre; Weiwei Guo; Rada Mihalcea; German Rigau; Janyce Wiebe. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. Proceedings of SemEval 2014.

10. Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, WeiWei Guo. *SEM 2013 shared task: Semantic Textual Similarity, Proceedings of *SEM 2013.

11. Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. Proceedings of SemEval 2012.

12. Bakhteev, O., Kuznetsova, R., Romanov, A. & Khritankov, A. (2015, November). 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), 2015 (pp. 3-10). IEEE.

13. Khritankov, A. S., Botov, P. V., Surovenko, N. S., Tsarkov, S. V., Viuchnov, D. V., & Chekhovich, Y. V. Discovering Text Reuse in Large Collections of Documents: a Study of Theses in History Sciences.

14. Kutuzov, A. (2014). Semantic clustering of Russian web search results: possibilities and problems. In Information Retrieval (pp. 320-331). Springer International Publishing.

15. Pronoza, E., & Yagunova, E. (2015). 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), 2015 (pp. 74-82). IEEE.

16. Pronoza, E., & Yagunova, E. (2015). Low-Level Features for Paraphrase Identification. In Advances in Artificial Intelligence and Soft Computing (pp. 59-71). Springer International Publishing.

17. Panchenko, A., Loukachevitch, N. V., Ustalov, D., Paperno, D., Meyer, C. M. and Konstantinova, N., (2015) RUSSE: The First Workshop on Russian Semantic Similarity, Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”, vol. 2, pp. 89-105, RGGU

18. 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

## A Training and testing file formats

Training data are available at http://www.paraphraser.ru/download/get?file_id=1

Testing data will be available in XML format, similar to that of the training data. In Fig. 1 there is an example of testing file.

```
<?xml version='1.0' encoding='UTF8' ?>
<data>
  <corpus>
    <paraphrase>
      <value name="id">1</value>
      <value name="id_1">1</value>
      <value name="id_2">2</value>
      <value name="text_1">sentence_1</value>
      <value name="text_2">sentence_2</value>
    </paraphrase>
    <paraphrase>
      <value name="id">2</value>
      <value name="id_1">3</value>
      <value name="id_2">4</value>
      <value name="text_1">sentence_3</value>
      <value name="text_2">sentence_4</value>
    </paraphrase>
    <paraphrase>
      <value name="id">2</value>
      <value name="id_1">5</value>
      <value name="id_2">6</value>
      <value name="text_1">sentence_5</value>
      <value name="text_2">sentence_6</value>
    </paraphrase>
    <!--    rest of the sentences   -->
  </corpus>
</data>
```

Fig. 1. Testing file format

## B Submission of results

### File naming conventions

There are no restrictions as to the names of the files.

### File formats

Files for both runs should be submitted in XML format, as in the example in Fig. 2.

```xml
<?xml version='1.0' encoding='UTF8'?>
<data>
  <team>abc123</team>
  <run type="Standard" no="5">
  <task>1</task>
  <response>
    <paraphrase>
      <value name="id">1</value>
      <value name="id_1">1</value>
      <value name="id_2">2</value>
      <value name="text_1">sentence_1</value>
      <value name="text_2">sentence_2</value>
      <value name="class">0</value>
    </paraphrase>
    <paraphrase>
      <value name="id">2</value>
      <value name="id_1">3</value>
      <value name="id_2">4</value>
      <value name="text_1">sentence_3</value>
      <value name="text_2">sentence_4</value>
      <value name="class">1</value>
    </paraphrase>
    <paraphrase>
      <value name="id">2</value>
      <value name="id_1">5</value>
      <value name="id_2">6</value>
      <value name="text_1">sentence_5</value>
      <value name="text_2">sentence_6</value>
      <value name="class">-1</value>
    </paraphrase>
    <!--    rest of the sentences   -->
  </response>
</data>
```

Fig. 2. Submission file format

In tag “task” participants should insert the number of the task. Tag “team” should contain the code of the team, and in tag “run” type should equal “Standard” or “Non-standard”, and “no” stands for the number of the submitted run.

For task 1, in tag “value” with “name” property equal to “class”, 1 stands for precise paraphrases, 0 stands for near paraphrases and -1 stands for non-paraphrases. For task 2, where there are only two classes, 1 stands for paraphrases (precise or near) and 0 stands for non-paraphrases.

**Data encoding**

UTF-8 without BOM.