Construction of a Russian Paraphrase Corpus: Unsupervised Paraphrase Extraction

Ekaterina Pronoza(✉), Elena Yagunova, and Anton Pronoza

Saint-Petersburg State University, Saint-Petersburg, Russian Federation
katpronoza@gmail.com, iagounova.elena@gmail.com, antpro@list.ru

Abstract. This paper presents a crowdsourcing project on the creation of a publicly available corpus of sentential paraphrases for Russian. Collected from the news headlines, such corpus could be applied for information extraction and text summarization. We collect news headlines from different agencies in real-time; paraphrase candidates are extracted from the headlines using an unsupervised matrix similarity metric. We provide user-friendly online interface for crowdsourced annotation which is available at paraphraser.ru. There are 5181 annotated sentence pairs at the moment, with 4758 of them included in the corpus. The annotation process is going on and the current version of the corpus is freely available at http://paraphraser.ru.

Keywords: Russian paraphrase corpus · Lexical similarity metric · Unsupervised paraphrase extraction · Crowdsourcing

1 Introduction

Our aim is to create a publicly available Russian paraphrase corpus which could be applied for information extraction (IE), text summarization (TS) and compression. We believe that such corpus can be helpful for paraphrase identification and generation for Russian and that is why we focus on the sentential paraphrases. Indeed, a sentential corpus does not impose any specific methods of further paraphrase identification or generation on the researcher. If such corpus is representative enough, it can serve as a dataset for the experiments on the extraction of word-, phrase- and syntactic level paraphrases.

Paraphrase is restatement of a text: it conveys the same meaning in another form. Such natural language processing (NLP) tasks as paraphrase identification and generation have been shown to be helpful for IE [25], question answering [14], machine translation [7], TS [19], text simplification [29], etc. Paraphrase identification is used to detect plagiarism [6] and to remove redundancies in TS [19] and IE [25], while paraphrase generation – to expand queries in information retrieval and question answering [14] and patterns – in IE. Paraphrase generation is also useful for text normalization [28] and textual entailment recognition tasks [9].

As far as the definition of paraphrase is concerned, it generally implies that the same message is expressed in different words, but it does not prescribe which portion of text is replaced by paraphrasing. Neither does it state whether common knowledge can be

P. Braslavski et al. (Eds.): RuSSIR 2015, CCIS 573, pp. 146–157, 2016.
DOI: 10.1007/978-3-319-41718-9_8

used when judging on the similarity of the two messages. As a consequence of this ambiguity, some researchers believe that paraphrases should have absolute semantic equivalence while others allow for bidirectional textual entailment, when the two messages convey roughly the same meaning.

Let us consider an example from our corpus: (1) ВТБ может продать долю в Tele2 в ближайшие недели. /VTB might sell its shares in TELE2 in the nearest weeks/ (2) ВТБ анонсировал продажу Tele2. /VTB announced the sale of TELE2/

Although it is clear that the two sentences describe the same event, the first one has additional details: indication of the time and the fact that the shares are going to be sold. A human judger, with his/her knowledge about the world, might consider these sentences paraphrases. But if we intend to teach a machine to identify semantically equivalent paraphrases, a threshold for paraphrases should be higher. On the other hand, the second sentence can be considered a summarization of the first one, and therefore such types of paraphrases can be used in automatic TS.

In our research, we intend to construct paraphrase corpus for IE and TS. We believe that the former task requires semantically equivalent, or precise paraphrases while the latter one demands roughly similar ones (so-called loose paraphrases) like those in our example. Thus, it is important for us to distinguish precise paraphrases (PP) and loose paraphrases (LP) while constructing our paraphrase corpus.

Today there are already a number of available paraphrase resources, Microsoft Paraphrase Corpus being the most well-known of them [13]. A wide number of metrics for paraphrase identification (for English) are evaluated against this corpus.

For Russian there are no publicly available paraphrase resources known to us, with the only exception of the dataset published by Ganitkevich et al. as part of The Paraphrase Database project [17]. The latter includes paraphrases on the word-, phrase- and syntactic levels, and each paraphrase pair is annotated with the set of count- and probability-based features. Such corpus can be used for both IE and TS, but it lacks information on the context of paraphrases. We believe that if such context (the original sentences) was provided, it could improve both these NLP tasks. That is why we aim at constructing a sentential corpus.

Thus, our task is to construct a corpus with both PPs and LPs and to make it helpful for paraphrase identification and generation in IE, TS and text compression tasks.

Our research is a part of an ongoing crowdsourcing project available at paraphraser.ru, with our current results available at paraphraser.ru/scorer/stat.

2 Related Work

In paraphrase identification/generation, unlike many other NLP applications, the data is usually hard to get. Paraphrases do not emerge naturally, like users' clicks or query logs, and gathering them manually is a tedious task. Moreover, one usually needs large amounts of data to collect just a few paraphrases.

Paraphrase corpora can be constructed from "natural" or "artificial" sources. The former usually include parallel multilingual corpora [2] and comparable monolingual corpora [22] (different translations of the same texts [11]; news texts/clusters [1, 10, 13,

27]; texts on similar topics, e.g., from the social networks (e.g., Twitter Paraphrase Corpus) [28] or students' answers to the questions [9]; social media [3], Wikipedia [26]; different descriptions of the same videos [8]), etc. "Artificial" sources are texts paraphrased by humans [20, 21, 24].

Gold standard paraphrases are typically extracted from the candidates set using either experts' [13, 20] or crowdsourced [1, 6, 8] annotation. 2-way and 3-way annotation is a common approach, but sometimes a complex system of characteristics is introduced (e.g., in [20] paraphrases are annotated along 10 dimensions of paraphrase characteristics on a 6 point scale).

A detailed overview of all the existing paraphrase corpora is beyond the scope of this paper. A thorough and insightful review of different sentential paraphrase datasets can be found in [21] where the authors present recommendations on paraphrase corpora construction and raise a number of important problems to the community.

Due to the aim of our research and space limitations we inevitably focus on the well-known Microsoft Research Paraphrase Corpus (MSRP) and on The Paraphrase Database, the only publicly available resource of Russian paraphrases known to us.

MSRP is not the oldest paraphrase corpus, but is definitely the one which greatly inspired research in paraphrase community. It was constructed as a broad-domain corpus of sentential paraphrases which would be amenable to statistical machine translation (SMT) techniques [13]. It consists of 5801 pairs of English sentences collected from news clusters and annotated by 2 experts. An initial set of paraphrases is extracted using Levenshtein edit distance. The authors only consider first 3 sentences of the articles and apply several criteria to their length and lexical distance between the sentences. The resulting dataset is extracted using SVM with morphological, lexical, string similarity and composite features.

Although MSRP is widely used as the gold standard in the experiments on paraphrase extraction methods, it is often criticized by researchers for its loose definition of paraphrase, for its 2-way annotation, high lexical overlap, etc.

While constructing a Russian corpus, we try to solve the problem of paraphrase ambiguity by distinguishing 2 types of paraphrases: precise and loose ones. We have 3-way annotation: precise paraphrases, loose paraphrases and non-paraphrases. As for the lexical overlap problem, we consider this overlap acceptable and even helpful in our case. Russian is a language with free word order, and pairs of sentences which consist of the same words put in different order could be used for learning syntactic patterns for paraphrase generation.

As we have already mentioned, there is one publicly available Russian paraphrase resource known to us: the dataset published by Ganitkevich et al. as part of The Paraphrase Database project (PPDB) [17]. The authors collected an impressively large database of paraphrases on word-, phrase- and syntactic levels. Syntactic level paraphrases are annotated with nonterminal symbols (constituents, in terms of phrase structure grammar) and contain placeholders which can be substituted with any paraphrase matching its syntactic type. In addition, all types of paraphrases are annotated with count and probability-based features. These features include the difference in the number of words/characters/average word length between the original phrase and the paraphrase, the probability of the original phrase given the paraphrase, alignment features, etc.

Some features are derived from the syntactic rules, e.g., the probability of the lefthand side nonterminal symbol given the paraphrase (and vice versa).

The training data for Russian is substantial in PPDB (over 2 million sentence pairs), and the resulting dataset is large as well. It is collected from the corpora typically used in SMT: CommonCrawl, Yandex 1M corpus and News Commentary. 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. Such approach was introduced by Bannard and Callison-Burch in 2005 [2] and since then it has been successfully applied by many researchers. The authors acknowledge that in morphologically rich languages different forms of the same word tend to group into the same paraphrase clusters because English phrases are chosen as the pivot ones (in Russian different forms of the same word are considered paraphrases in PPDB). While for some tasks it could be desirable, for others it is definitely not (it could cause generation of incorrect paraphrases). Moreover, such grouping leads to the rapid growth of the dataset, and, with a number of available morphological parsers today, it seems unnecessary. We also believe that we should use language-specific methods (in contrast with language-independent ones) when dealing with a morphologically rich language.

Unlike other paraphrase resources, our corpus is not intended to be a general-purpose one. According to our tasks (IE and TS) we collect it from the news texts. The corpus consists of sentential paraphrases, and lower level paraphrase pairs can be extracted from it using any of the existing methods (e.g., SMT methods).

3 Unsupervised Paraphrase Extraction

3.1 Data: Method

We adopt a sentence-level approach and extract paraphrases from the news articles published on the Web. The latter is a truly rich source of paraphrases: the articles describing the same events appear in different newspapers every day.

Due to the lack of training data for Russian, our approach is unsupervised. It extends the one described by Fernando and Stevenson [15]: their method yields the best results against MSRP among the latest unsupervised approaches. Our hypothesis is that paraphrases can be successfully extracted from the Russian news texts based on a lexical similarity metric.

We automatically extract articles published by several newspapers on the same day during the last 2 years. Then we adopt the strategy by Wubben et al. [27] and proceed with pairwise comparisons of headlines. A headline of an article can be considered its compression[1], and we suppose that the headlines of the articles describing the same events are similar and may even be paraphrases of each other. Moreover, headlines

[1] This statement can only be applied to the informative news texts (the ones intended to inform, and not to persuade the reader) and not to the publicistic texts (exerting influence on the reader in the first place). A publicistic headline is often designed to attract readers' attention. However, both publicistic and informative texts can be used as a source of paraphrases.

comparison is much faster than the comparison of all the sentences from the two articles. We do not take into account too short headlines (less than 3 words long) as they are unlikely to add to the representativeness of the resulting corpus.

The overall scheme is as follows: we iterate over all possible pairs of headlines (with the same date of publication) from different media agencies and calculate a similarity metric for each pair. Then the pairs with scores below the threshold value are pruned with the exception of a small portion of the necessary negative instances. The resulting dataset is evaluated by the annotators. Having the annotated data, we further adopt a supervised approach and optimize the unsupervised similarity metric.

3.2 Sentence Similarity Metric

To extract paraphrases, we use an unsupervised lexical similarity metric based on the one proposed by Fernando and Stevenson [15]:

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a}W\vec{b}}{\left|\vec{a}\right|\left|\vec{b}\right|}, \tag{1}$$

where W is a similarity matrix and $\vec{a}$ and $\vec{b}$ – word vectors of the two sentences. Each element w_{ij} of the matrix represents the similarity between the words a_i and b_j. Diagonal elements obviously equal 1. Other elements equal 0 for different or $0 < \omega < 1$ for similar words. To capture lexical similarity, the authors use several metrics mainly based on the "*is-a*" hierarchy from WordNet [15].

In our research we use a matrix metric with lexical similarity scores based on the synonymy relation. As far as the source of synonymy relations is concerned, there exists a famous Russian dictionary of synonyms by Abramov [30] created over a century ago. Despite its numerous merits, the dictionary is deplorably outdated. The lack of modern synonymy resources has spurred a number of attempts at creating databases of synonyms. Although most resources are designed for practical purposes (rewriting texts in the web and automatically producing unique content), they can also be useful for NLP in general, and for our task in particular. We use one of such collections which consists of about 6 thousand articles. In fact, each article (a word and the list of its synonyms) can be considered a synset. For every pair of words we may further calculate the number of times they occur together in the same synset. In terms of information retrieval, every synset is a document, and it is known that semantically close are more likely to appear in the same documents than in the different ones. To compute lexical similarity, we use such metrics as normalized pointwise Mutual Information (npmi) [4], Dice coefficient [12] and Jaccard index [18].

Unlike the original metric from [15] (which uses WordNet relations), ours is calculated according to the list of scoring rules designed to capture not only synonymy relations but also conjugate words[2]:

[2] The latter might be of no importance for English, but they are essential for detecting Russian sentential paraphrases.

– Identical words starting with capital letters -> 1.2 score (a slight bias towards the simultaneous occurrence of the same named entities in the sentences).
– Identical words -> 1.
– Synonyms -> Npmi, Dice or Jaccard coefficient multiplied by 0.8.
– One of the words is a substring of the other -> the score equal to the length of the smaller word divided by the length of the larger word and multiplied by 0.7.
– The words have common prefix (at least 3 characters) -> the score equal to the prefix length divided by the length of the lesser word and multiplied by 0.
– Otherwise -> 0.

The original metric varies from 0 to 1. With our modifications it no longer satisfies this condition but it does not affect paraphrase extraction process. The scores (1.2, 1, 0.8, etc.) are obtained from the preliminary experiments conducted on the small subset of the corpus. Based on the results of these experiments, we select Jaccard index as the synonymy coefficient in our metric.

Let us calculate the metric for the two sentences from our dataset:

1. КНДР аннулировала договор о ненападении с Южной Кореей. /DPRK annulled the non-aggression treaty with South Korea/
2. КНДР вышла из соглашений о ненападении с Южной Кореей. /DPRK withdrew from the non-aggression agreement with South Korea/

We lemmatize the sentences using TreeTagger [23] and cut off auxiliary words. After these manipulations we represent the sentences as binary vectors (Fig. 1):

Word	аннулировать /annul/	выйти /withdraw/	договор /treaty/	КНДР /DPRK/	Корея /Korea/	ненападение /non-aggression/	соглашение /agreement/	Южный /South/
S1 =	(1,	0,	1,	1,	1,	1,	0,	1)
S2 =	(0,	1,	0,	1,	1,	1,	1,	1)

Fig. 1. Example of word vectors for the two sentences

One can see that there are 4 overlapping words, 3 of them starting with an uppercase letter, and 2 synonyms: “соглашение” (agreement) and “договор” (treaty). The word “соглашение” occurs in 5 synsets in the dictionary of synonyms, while “договор” – only once (and their appear in one synset together). Jaccard index equals $1/(1 + 5 - 1) = 0.2$ for the two given words. This score is multiplied by the pruning coefficient: $0.2 * 0.8 = 0.16$. The similarity matrix for the two sentences is shown in Fig. 2.

According to (1), the resulting similarity score equals 0.763.

We apply the described metric to the pairs of article headlines and prune the ones with the Jaccard index-based similarity score below the empirically defined threshold value of 0.5.

Thus, our approach is based on the existing similarity metric, but according to our goal – the construction of paraphrase corpus for IE and TS – we introduce a list of scoring

	аннулировать	выйти	договор	КНДР	Корея	ненападение	соглашение	Южный
аннулировать	1	0	0	0	0	0	0	0
выйти	0	1	0	0	0	0	0	0
договор	0	0	1	0	0	0	0.16	0
КНДР	0	0	0	1.2	0	0	0	0
Корея	0	0	0	0	1.2	0	0	0
ненападение	0	0	0	0	0	1	0	0
соглашение	0	0	0.16	0	0	0	1	0
Южный	0	0	0	0	0	0	0	1.2

Fig. 2. Similarity matrix example

rules which capture the linguistic phenomena (synonymy relations, conjugate words, matching names entities) required for our corpus.

4 Corpus Annotation and Analysis

4.1 Annotation

Potential paraphrases with similarity metric values above the threshold are evaluated by the annotators. To obtain negative instances, we also include a portion of random sentence pairs with metric value below 0.5 in the corpus (roughly speaking, every fourth sentence pair in the corpus has a score below 0.5).

At the moment there are 5424 pairs of sentences in the corpus, with 5181 annotated pairs. Out of these 5181 pairs of sentences, we select 4758 pairs and include them in the corpus (by pruning inconsistent or potentially unreliable results).

We developed an online interface: http://paraphraser.ru/scorer for crowdsourced annotation. A user is shown two sentences at a time, and he/she is to decide whether the sentences convey the same meaning (class "1"), similar meanings (class "0") or different meanings (class "−1"). The users are advised to user their own judgement and intuition and are not given any specific instructions.

We try to make the annotation process less tedious by introducing an entertainment element: the users are shown various facts (about different events like the invention of something or the birth of a famous scientist/artist, etc.) and pictures at random intervals and are encouraged to annotate further.

It is well known that crowdsourcing poses a challenge concerning the reliability of the obtained results. To prune unreliable results, we only consider sentence pairs annotated by at least 3 users. If a paraphrase pair is annotated by less than 4 users, and two of them provide opposite judgments ("−1" and "1"), such pair is cut off. In future we also plan to involve expert linguists in the annotation process.

To assign a class to each sentence pair ("−1" for non-paraphrases, "0" for LPs and "1" – for PPs), we compute the median of all the scores given to the pair by the annotators. It can obviously take one of the following values: {− 1, −0.5, 0, 0.5, 1}. As we would like to have only 3 classes, in case of ties we adopt a pessimistic strategy and round the value down to the previous integer (−0.5 is reduced to −1, 0.5 – to 0).

4.2 Paraphrase Classes

As stated earlier, we distinguish non-paraphrases, loose paraphrases (LPs) and precise paraphrases (PPs). Their distribution in our corpus is presented in Table 1.

Table 1. Distribution of paraphrase classes in the corpus

Paraphrase class	Number of instances	Percentage of instances
Non-paraphrases	1599	33.6 %
Loose paraphrases	1969	41.4 %
Precise paraphrases	1190	25 %
Total	*4758*	*100 %*

Although one cannot say that the dataset is severely unbalanced, there is a slight bias towards loose paraphrases. To analyze the differences between precise and loose paraphrases in the corpus we follow the approach adopted in [13]: we randomly select 100 PPs and 100 LPs and manually annotate them with the linguistic features:

- different content (sentences differ in words or phrases which carry additional information and make the sentences semantically different);
- different time (different grammar tenses are used to described the same event);
- context knowledge (sentences differ in the words or phrases, and added words/phrases have no counterparts in the other sentences (see "different content"), but nevertheless it is clear from the context of the phrases that the same notion/event is being referred to, and that there is no semantic difference);
- metaphor (a metaphor takes place in one of the sentences);
- metonymy (sentences differ in some named entities, and one of these entities is used metonymically);
- numeral (sentence pairs differ in the representation of the same numerals or the number is rounded in one of the sentences);
- phrasal synonymy (sentences differ in the synonymous multiword expressions);
- reordering (sentences consist of the same words in different order);
- word-level synonymy (sentence pairs differ in the synonymous words);
- syntactic synonymy (the same information is expressed in the sentences using different constituents or the same constituents with different grammatical characteristics, e.g., a verbal phrase in active and passive voice respectfully).

Each pair of sentences can be annotated with more than one linguistic feature (e.g., syntactic synonymy is often accompanied by reordering and word-level synonymy). For each of the features we calculate the portion of sentence pairs it occurs in (see Table 2). These portions are calculated for PPs and LPs separately.

It can be seen that metaphor, metonymy and different representations of numbers are rare events in both types of paraphrases in our sample. While most (76 %) LPs differ in the meaning they convey (see "different content"), PPs are richer in word-, phrase- and syntactic-level synonymy. In fact, such results are quite predictable, and it is just what we expect these two paraphrase classes to be like. However, the portion of different content (18 %) among PPs (this feature is undesirable for PPs in our corpus) is not what

one could call neglectable. Indeed, deciding on the semantic equivalence is a challenging task even for linguist experts, setting aside mere native speakers. Thus, in future we plan to involve experts' annotation to reduce the portion of semantically different phrases among PPs.

Table 2. Linguistic characteristics of precise and loose paraphrases

Feature	PP	LP	Feature	PP	LP
Context knowledge	32 %	25 %	Numeral	9 %	4 %
Different content	18 %	76 %	Phrasal synonymy	16 %	6 %
Different time	6 %	15 %	Reordering	17 %	10 %
Metaphor	0 %	1 %	Word-level synonymy	18 %	7 %
Metonymy	3 %	2 %	Syntactic synonymy	33 %	17 %

5 Evaluation

We evaluate the unsupervised metric used in the corpus construction by comparing it with the annotation results (see Table 3).

Table 3. Unsupervised paraphrase extraction: results

	Score above threshold	Score below threshold	Total
Precise paraphrases	1179	11	1190
Loose paraphrases	1919	50	1969
Non-paraphrases	763	836	1599
Total	3861	897	4758

As we do not distinguish between PPs and LPs when collecting sentence pairs using Jaccard-based metric, we only evaluate the quality of the metric in the task of classifying sentence pairs into similar (PPs and LPs) and different ones. Thus, its precision equals 80.24 %. We believe that the evaluation via traditional recall and F1 measures would be unreliable in our case. We do not focus on the collection of a balanced dataset with "proper" negative instances at the moment (approximately every fourth candidate is randomly selected as a potential negative instance), and recall and F1-score would overestimate our metric.

Our unsupervised metric extends the one used by Fernando and Stevenson [15]. They evaluated it against MSRP and achieved 75.2 % precision. Thus, our result seems promising although one should bear in mind that it only reflects the quality of classifying sentence pairs into 2 classes, when PPs and LPs are merged into one class.

Having obtained the annotated data, we optimized our metric: the threshold and the scores changed, Dice coefficient was selected instead of Jaccard index, and the overall performance improved, but due to space limitations we cannot give full details of the experiments here. At the moment we are working on a supervised approach towards paraphrase identification and train a classifier to distinguish between PPs and LPs, with the optimized similarity metric being used as one of the features. In future we intend to

develop an approach which focuses on covering various paraphrase classes and linguistic phenomena [16, 21] because Russian is rich in these phenomena.

6 Conclusion: Future Work

In this paper we presented our work on the creation of a Russian sentential paraphrase corpus. The corpus consists of news headlines automatically collected from the Web and filtered using the unsupervised similarity metric. Such resource can be used in information extraction and text summarization. It can also serve as a training dataset for paraphrase identification models for Russian.

There are 4758 sentence pairs in the corpus at the moment, and it is freely available at our website: paraphraser.ru. All the pairs are being annotated using crowdsourcing via the website, and one of the three classes (non-paraphrase, loose paraphrase or precise paraphrase) is assigned to each pair of sentences. To obtain reliable data, we ensure that each pair of sentences in the corpus is annotated by at least 3 users and cut off inconsistent annotation. Evaluated against crowdsourced annotation, the similarity metric achieves 80.24 % precision at classifying paraphrases. Thus, it confirms our hypothesis that paraphrases can be extracted from the Russian news texts using methods based on lexical similarity.

Our further step aims at the development of paraphrase identification model and we are already working on it. This step includes using a better synonymy resource: Yet Another RussNet [5] (it is 8 times larger than our original one), and a dictionary of word formation families. We already use the optimized similarity metric in the paraphrase classifier and experiment with features based on semantic distributional models; other features are derived from the morphological characteristics, syntactic and semantic structure of the sentences. Thus, we intend to develop a fine-grained approach towards identifying paraphrases. As it might demand experts' annotation, it is one of our future work directions, along with the comparison of experts' annotation and the results of the automatic extraction of linguistic features. We acknowledge Saint-Petersburg State University for the research grant 30.38.305.2014.

References

1. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo W.: SEM 2013 shared task: semantic textual similarity. In: The Second Joint Conference on Lexical and Computational Semantics (2013)
2. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of the 43rd Annual Meeting of the ACL, pp. 597–604 (2005)
3. Bernhard, D., Gurevych, I.: Answering learners' questions by retrieving question paraphrases from social Q&A sites. In: Proceedings of the ACL 2008 3rd Workshop on Innovative Use of NLP for Building Educational Applications, pp. 44–52 (2008)
4. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of the Biennial GSCL Conference (2009)

5. Braslavski, P., Ustalov, D., Mukhin, M.: A spinning wheel for YARN: user interface for a crowdsourced thesaurus. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, pp. 101–104 (2014)
6. Burrows, S., Potthast, M., Stein, B.: Paraphrase acquisition via crowdsourcing and machine learning. ACM Trans. Intell. Syst. Technol. **4**(3), 43 (2013)
7. Callison-Burch, C.: Paraphrasing and Translation. Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh (2007)
8. 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, Portland, Oregon, USA, pp. 190–200 (2011)
9. Dzikovska, M.O., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., 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)
10. Clough, P., Gaizauskas, R., Piao, S., Wilks, Y.: METER: MEasuring TExt Reuse. In: Isabelle, P. (ed.) Proceedings of the Fortieth Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 152–159 (2002)
11. Cohn, T., Callison-Burch, C., Lapata, M.: Constructing corpora for the development and evaluation of paraphrase systems. Comput. Linguist. Arch. **34**(4), 597–614 (2008)
12. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
13. Dolan, W.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, Geneva, Switzerland (2004)
14. Duboue, P.A., Chu-Carroll, J.: Answering the question you wish they had asked: the impact of paraphrasing for question answering. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, New York, pp. 33–36 (2006)
15. Fernando, S., Stevenson, M.: A semantic similarity approach to paraphrase detection. In: Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloqium (2008)
16. Fujita, A., Inui, K.: A class-oriented approach to building a paraphrase corpus. In: Proceedings of the Third International Workshop on Paraphrasing (2005)
17. Ganitkevitch, J., Callison-Burch, C.: The multilingual paraphrase database. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). European Language Resources Association (ELRA), Reykjavik (2014)
18. Jaccard, P.: Étude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 547–579 (1901)
19. Knight, K., Marcu, D.: Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artif. Intell. **139**(1), 91–107 (2002)
20. McCarthy, Ph.M., McNamara, D.S.: The user-language paraphrase corpus. In: Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches, pp. 73–89 (2008)
21. Rus, V., Banjade, R., Lintean, M.: On paraphrase identification corpora. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 2422–2429. European Language Resources Association (ELRA), Reykjavik (2014)
22. Sanchez-Perez, M., Sidorov, G., Gelbukh, A.: The winning approach to text alignment for text reuse detection at PAN 2014. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings, vol. 1180, pp. 1004–1011. CEUR-WS.org (2014). ISSN: 1613-0073

23. Schmid, H.: Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland (1995)
24. Shimohata, M., Sumita, E., Matsumoto, Y.: Building a paraphrase corpus for speech translation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004). European Language Resources Association (ELRA), Lisbon (2004)
25. Shinyama, Y., Sekine, S.: Paraphrase acquisition for information extraction. In: Proceedings of the Second International Workshop on Paraphrasing, vol. 16, pp. 65–71 (2003)
26. Vila, M., Rodriguez, H., Marti, M.A.: WRPA: a system for relational paraphrase acquisition from wikipedia. Procesamiento del Lenguaje Nat. **45**, 11–19 (2010)
27. 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, Athens, Greece, pp. 122–125 (2009)
28. Xu, W., Ritter, A., Grishman, R.: Gathering and generating paraphrases from twitter with application to normalization. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, pp. 121–128 (2013)
29. Zhao, Sh., Lan, X., Liu, T., Li, Sh.: Application-driven statistical paraphrase generation. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, pp. 834–842 (2009)
30. Abramov, N.: Slovar' russkih synonymov I shodnyh po smislu virazheniy, 7th edn. Russkie slovari, Moscow (1999)