Opportunities and Challenges in Semantic Similarity

Vasile Rus
Department of Computer Science
The University of Memphis
Memphis, TN 38152
vrus@memphis.edu

Abstract

Semantic similarity has been increasingly adopted in the recent past as a viable, scalable alternative to the full-understanding approach to natural language understanding. We present here an overview of opportunities and challenges to semantic similarity with an emphasis on methods, data, and tools. A series of methods we developed over the past decade will be summarized. These methods and others have been integrated in a semantic similarity toolkit called SEMILAR (www.semanticsimilarity.org) which has been widely adopted by thousands of users since its launch in summer of 2013 at the Annual Meeting of the Association for Computational Linguistics. Furthermore, we will illustrate some drawbacks of current data sets that hamper a fair comparison among existing methods. Several suggestions will be made to improve the building of future data sets for assessing approaches to semantic similarity.

Introduction

The task of semantic similarity has become a mainstream research area in the wider natural language processing research community. It has become to be viewed as a viable, scalable, and cost-effective solution to the central issue of natural language understanding, i.e. the understanding of natural languages by machines. Indeed, semantic similarity has proven to be a robust alternative to the true understanding approach to natural language understanding. Because the true understanding approach is knowledge intensive it is less scalable and cost-prohibitive. For instance, a full understanding approach relies on word knowledge an intractable problem.

We define the problem of semantic similarity between two texts, denoted A and B, as quantifying the semantic relation between the two texts, e.g. to what extent text A has the same meaning as text B (paraphrase relation) or to what extent text A entails text B (entailment relation), and/or making a qualitative decision about the type of semantic relationship between the two texts. For the qualitative case, the task would be to decide, usually based on some quantified measures, whether the two texts are in a particular semantic relationship or not.

The importance of the semantic similarity task in Natural Language Processing (NLP) is highlighted by the diversity of datasets and shared task evaluation campaigns (STECs) that have been proposed over the last decade (Dolan, Quirk, and Brockett, 2004; McCarthy & McNamara, 2008; Agirre et al., 2012). For instance, the task of paraphrase identification, an instance of the semantic similarity problem, is important for a number of applications including Natural Language Generation, Question Answering, and dialogue-based Intelligent Tutoring Systems. In Natural Language Generation, paraphrases are a method to increase diversity of generated text (Iordanskaja et al. 1991). In Question Answering, multiple answers that are paraphrases of each other could be considered as evidence for the correctness of the answer (Ibrahim et al. 2003). In Intelligent Tutoring Systems (Graesser et al. 2005; McNamara et al. 2007), paraphrase identification is useful to assess whether student's articulated answers to deep questions (e.g. conceptual physics questions) are similar-to/paraphrases-of ideal answers.

As a concrete example of a semantic similarity task, we show below a pair of sentences from the Microsoft Research Paraphrase Corpus (Dolan, Quirk, and Brockett, 2004) in which Text A is a paraphrase of Text B and vice versa.

Text A: *York had no problem with MTA's insisting the decision to shift funds had been within its legal rights*.
Text B: *York had no problem with MTA's saying the decision to shift funds was within its powers*.

Given such two texts, the challenge is to automatically assess whether Text A is a paraphrase of Text B. Other types of semantic relations among two texts have been explored such as textual entailment (Dagan, Glickman, & Magnini, 2004), i.e. whether text A entails text B, or elaboration (McCarthy & McNamara, 2008), i.e. whether text B is an elaboration of text A. In general, approaches to the tasks of paraphrase identification, recognizing textual entailment, or elaboration detection first quantify along various dimensions how semantically similar the two texts are and then make a qualitative decision such as a paraphrase relation does exist between the two texts (see the binary qualitative decisions in MSRP) or more nuanced decisions are made (see the recent Semantic Textual Similarity task at SemEval; Agirre et al., 2012).

Semantic similarity can be broadly construed as being assessed between any two texts of any size. Depending on the granularity of the texts, we can talk about the following fundamental text-to-text similarity problems: word-to-word similarity, phrase-to-phrase similarity, sentence-to-sentence similarity, paragraph-to-paragraph similarity, or document-to-document similarity. Mixed combinations are also possible such as assessing the similarity of a word to a sentence or a sentence to a paragraph. For instance, in summarization it might be useful to assess how well a sentence summarizes an entire paragraph.

The rest of the paper is organized as in the followings. The next section provides an overview of related work followed by a description of the data, i.e. the corpus of speeches. Then, we describe the major methods we used to measure cohesion in Arabic (for English we use standard methods which can be found in the works cited in the Previous Work section) and the results obtained. We conclude the paper with Discussion and Conclusions.

Data Sets

While research on word-to-word similarity measures was conducted for more than a decade (Pedersen, Patwardhan, & Michelizzi, 2004), semantic similarity as a mainstream research areas spawned with the development and public release of the Microsoft Research Paraphrase corpus (Dolan, Quirk, and Brockett, 2004).

One of the most important legacies of the MSRP corpus is the inspiration it created for the development of other corpora for paraphrase research in particular and for other semantic relations such as elaboration. In fact, some of the more recent corpora use some of the earlier corpora as a source, e.g. Cohn, Callison-Burch, and Lapata (2008), use MSRP as a source for building their corpus; similarly, the STS challenge borrowed a portion of MSRP corpus. We will describe next the major data sets in the area of semantic similarity starting the MSRP.

The Microsoft Research Paraphrase corpus (MSRP; Dolan, Quirk, and Brockett, 2004) consists of 5,801 newswire sentence pairs, 3,900 of which were labeled as paraphrases by human annotators. The MSRP corpus is divided into a training set (4,076 sentence) which we have used to determine the optimum threshold, and a test set (1,725 pairs) that is used to report the performance scores. Average words per sentence number for this corpus is 17. MSRP is by far the largest publicly available paraphrase annotated corpus, and has been used extensively over the last decade.

ULPC the User Language Paraphrase Corpus (ULPC; McCarthy and McNamara 2008), which contains pairs of target-sentence/student response texts. These pairs have been evaluated by expert human raters along 10 dimensions of paraphrase characteristics. In current experiments we evaluate the LSA scoring system with the dimension called ”Paraphrase Quality bin”. This dimension measures the paraphrase quality between the target-sentence and the student response on a binary scale, similar to the scale used in MSRP. From a total of 1998 pairs, 1436 (71%) were classified by experts as being paraphrases. A quarter of the corpus is set aside as test data. The average words per sentence number is 15.

The Question Paraphrase corpus contains 1,000 questions along with their paraphrases (totaling 7,434 question paraphrases) from 100 randomly selected FAQ files in the Education category of the WikiAnswers web site (Bernhard & Gurevych, 2008). The 1,000 questions are called the target questions and the 7,434 question paraphrases are called the input questions. The objective of their paraphrase task is to retrieve the corresponding target question for each input question. That is, their corpus contains 7,434 true paraphrases or, from another perspective, their corpus contains 1000 target questions for which there are on average 7.434 paraphrased questions. There is no explicit representation of false paraphrase instances.

The SEMILAR, formerly known as SIMILAR, corpus (Rus et al., 2012) is the richest corpus in terms of annotated information and scope, e.g. it can be used for assessing word-to-word similarity measures, word-to-word similarity measures in context, sentence level paraphrase identification methods, and alignment algorithms. The SEMILAR corpus contains 700 pairs of sentences from the MSRP corpus: 29,771 tokens (words and punctuation) of which 26,120 are true words and 17,601 content words. The number of content words is important because many word-to-word semantic similarity metrics available work on content words or certain types of content words, e.g. only between nouns or between verbs. The 700 pairs are fairly balanced with respect to the original MSRP judgments, 49% (344/700) of the pairs are TRUE paraphrases. The corpus creators re-judged the semantic

equivalence of the selected instances. Their judgments yielded 63% (442) TRUE paraphrases for an overall agreement rate between their annotations and the MSRP annotations (both TRUE and FALSE paraphrases) of 75.7%. The judges were simply instructed to use their own judgment with respect to whether the two sentences mean the same thing or not. It should be noted that the MSRP guidelines were more targeted, e.g. judges were asked to consider different numerical values as being equivalent while we left such instructions unspecified. These differences in guidelines may explain the disagreements besides the personal differences in the annotators' background.

The SEMILAR corpus can be considered the richest in terms of annotation as besides holistic judgments of paraphrase they provide several word level similarity and alignment judgments. The corpus includes a total of 12,560 expert-annotated relations for a greedy word-matching procedure and 15,692 relations for an optimal alignment procedure.

The Student Response Analysis corpus (SRA; Dzikovska et al., 2013) consists of student answer-expert answer pairs collected from two intelligent tutoring systems. Both student answers and expert answers were related to specific tutorial questions from different science domains. There are 56 questions and 3,000 student answers from the so-called BEETLE corpus and 197 assessment questions and 10,000 answers from the ScientsBank corpus. These pairs were annotated using a combination of heuristics and manual annotation. They used a 5-way annotation as opposed to the typical 2-way annotation used in previous corpora.

The Semantic Textual Similarity corpus (STS; Agirre et al., 2013) contains 2,250 pairs of headlines, machine translation evaluation sentences, and glosses (concept definitions). The data set is balanced and they also used string similarity for selection of instances. We only describe here the STS CORE corpus as its input is pure task. The additional STS TYPE corpus provided metadata which makes it a bit different from a typical sentience-level paraphrase task. The STS CORE corpus was annotated through crowdsourcing. The annotation used a 6-way schema ranging from 5=identical to 0=completely unrelated. An earlier version of the corpus was used in 2012 for a pilot STS challenge. The training data contained 2000 sentence pairs from previously existing paraphrase datasets and machine translation evaluation resources. The test data also comprised 2000 sentences pairs for those datasets, plus two surprise datasets with 400 pairs from a different machine translation evaluation corpus and 750 pairs from a lexical resource mapping exercise. The similarity of pairs of sentences was rated on a 0-5 scale (low to high similarity) by human judges using Amazon Mechanical Turk.

Regneri and Wang (2012) built a dataset starting with 2000 sentence pairs collected from recaps of episodes of the TV show House, M.D. Among all gold standard sentence pairs, they found 158 paraphrases, 238 containment cases, 194 related pairs, and 1,402 unrelated. After discarding 8 sentence pairs and collapsing the categories of paraphrase, containment, and related, they ended up with 27% of the 590 instances in a broader paraphrase category (proper paraphrases) and 73% of them containing additional information that does not belong to the paraphrased part.

Rus & Graesser (2006) presented a small data set of sentence-level paraphrases collected from an intelligent tutoring environment. One expert physicist rated the degree to which particular speech acts expressed during training with a computer tutor matched particular expectations. These judgments were made on a sample of 25 physics expectations (E) and 5 randomly sampled student answers (S) per expectation, yielding a total of 125 pairs of expressions. The learner answers were always responses to the first hint for that expectation. The E-S pairs were graded by Physics experts on a scale of 1-4 (4 being perfect answer). This rubric could be used to prepare entailment tasks that deliver not only TRUE-FALSE decisions but also more fine-grained outputs. However, we followed the current RTE guidelines and transformed these numerical values to a discrete metric: scores 3 and 4 equal a TRUE decision and 1 and 2 equal a FALSE decision. We ended up with 36 FALSE and 89 TRUE entailment pairs, i.e. a 28.8% versus 71.2% split (as compared to the 50-50% split of RTE data).

Cohn, Callison-Burch, and Lapata (2008) a started with the Multiple-Chinese Translation corpus, Jules Vernes' Twenty Thousands Leagues Under the Sea novel, and MSRP. They selected 300 pairs of sentences from each source for a total 900 instances. They first asked annotators to align the sentences in each pair at word level. The authors argue that their corpus can be used for analyzing structural paraphrases besides lexical paraphrases.

Other data sets for paraphrase exist but they do not fit in the general category of sentence-level paraphrases, e.g. Potthast and colleagues (2010) created the PAN corpus which contains paragraph-size texts, Lintean, Rus, and Azevedo (2011) describe another paragraph-level paraphrase corpus, and Rodney and colleagues (2008) who created a sentence level data set for textual entailment in the context of science learning with an intelligent tutoring system.

As can be noted, there are a quite large number of data sets available that vary in the source of text, size, and annotations. The lack of annotation consistency or compatibility among some of these data sets makes a direct comparison of various approaches using different data sets less meaningful. One of reasons that explains the

differently proposed annotation schemes is the inherent difficulty of defining what a paraphrase is as noted by us in a recent paper (Rus, Banjade, & Lintean, 2014).

Methods

A myriad of methods have been proposed during the last decade or so to address the task of paraphrase identification or related tasks. It is beyond the scope of this paper to offer a comprehensive list of the methods. Instead, we will discuss a series of related methods we have been exploring and which are embedded in our SEMILAR toolkit (Rus et al., 2014). All the methods below rely more or less on the compositionality principle: the meaning of a sentence is the results of the meaning of its individual words and the way they combine, in our case through syntactic relations. We will start with the simplest method, lexical overlap, and end with a newly proposed method that outperforms other known methods.

Lexical Overlap. Given two texts, the simplest method to assess their semantic similarity is to compute lexical overlap, i.e. how many words they have in common. That is, one counts the common words and they divide by a normalization factor. When counting the common words, should one lemmatize the words first or not? What the normalization factor be (the average length of the two texts or the longest text?) These questions suggests that that there are many lexical overlap variations. Indeed, a closer look at lexical overlap reveals a number of parameters that turns the simple lexical overlap problem into a large space of possibilities. Thousands of variants of lexical overlap can be generated by different parameter settings. Importantly, performance of these methods on paraphrase identification and textual entailment tasks can vary widely (Rus, Banjade, & Lintean, 2014). Some lexical overlap variations lead to performance results rivaling more sophisticated, state-of-the-art methods (Lintea, 2011; Rus, Banjade, & Lintean, 2014).

We present next a set of text-to-text similarity methods that rely on word-to-word (*w2w*) similarity measures. That is, these methods compute the similarity of larger texts using individual word similarities. These methods assume that a *w2w* is available through some external procedure such as Latent Semantic Analysis or WordNet-based similarity measures or the more recently proposed *w2w* based on Latent Dirichlet Allocation (Rus, Niraula, & Banjade, 2013).

Rus and Lintean (2012; Rus-Optimal Matching or ROM) proposed an *optimal* solution for text-to-text similarity based on word-to-word similarity measures. The optimal lexical matching is based on the optimal assignment problem, a fundamental combinatorial optimization problem which consists of finding a maximum weight matching in a weighted bipartite graph.

Given a weighted complete bipartite graph $G = X \cup Y; X \times Y)$, where edge xy has weight $w(xy)$, the optimal assignment problem is to find a matching M from X to Y with maximum weight.

A typical application is about assigning a group of workers, e.g. words in text A in our case, to a set of jobs (words in text B in our case) based on the expertise level, measured by $w(xy)$, of each worker at each job. By adding dummy workers or jobs we may assume that X and Y have the same size, n, and can viewed as $X = \{x_1, x_2, \dots, x_n\}$ and Y = $\{y_1, y_2, \dots, y_n\}$. In the semantic similarity case, the weight $w(xy)$ is the word-to-word similarity between a word x in text A and a word y in text B.

The assignment problem can also be stated as finding a permutation π of {1, 2, 3, … , n} for which $\sum_{i=1}^{n} w(x_i y_{\pi(i)})$ is maximum. Such an assignment is called optimum assignment. An algorithm, the Kuhn-Munkres method (Kuhn, 1955), has been proposed that can find a solution to the optimum assignment problem in polynomial time.

Rus and colleagues (Rus et al., 2009; Rus & Graesser, 2006; Rus-Syntax-Negation or RSN) used a lexical overlap component combined with syntactic overlap and negation handling to compute an unidirectional subsumption score between two sentences, T (Text) and H (Hypothesis). Each text is regarded as a graph with words as nodes and syntactic dependencies as edges. The subsumption score reflects how much a text is subsumed or contained by another. The equation below provides the overall subsumption score, which can be averaged both ways to compute a similarity score, as opposed to just the subsumption score, between the two texts.

$$entscore(T,H) = (\alpha \times \frac{\sum_{V_h \in H_v} \max_{V_t \in T_v} match(V_h, V_t)}{|V_h|} + \beta \times \frac{\sum_{E_h \in H_e} \max_{E_t \in T_e} match(E_h, E_t)}{|E_h|}) \times \frac{(1+(-1)^{\#neg\_rel})}{2}$$

The lexical component can be used by itself (given a weight of 1 with the syntactic component given a weight of 0) in which case the similarity between the two texts is just an extension of *w2w* similarity measures. The *match* function can be any *w2w* similarity measure.

Lintean and Rus (2010; weighted-LSA or *w*LSA) extensively studied methods for semantic similarity based on Latent Semantic Analysis (LSA; Landauer et al., 2006). LSA represents words as vectors in a 300-500 dimensional LSA space. An LSA vector for larger texts can be derived by vector algebra, e.g. by summing up the individual words' vectors. The similarity of two texts A and B can be computed using the cosine (normalized dot product) of their LSA vectors. Alternatively, the individual word vectors can be combined through weighted sums. Lintean

and Rus (2010) experimented with a combination of 3 local weights and 3 global weights. All these version of LSA-based text-to-text semantic similarity measures are available in SEMILAR.

We also implemented a set of similarity measures based on the unsupervised method **Latent Dirichlet Allocation (LDA**; Blei, Ng, & Jordan, 2003**).** LDA is a probabilistic generative model in which documents are viewed as distributions over a set of topics (θ_d text d's distribution over topics) and topics are distributions over words (φ_t – topic t's distribution over words). That is, each word in a document is generated from a distribution over words that is specific to each topic.

A first LDA-based semantic similarity measure among words would then be defined as a dot-product between the corresponding vectors representing the contributions of each word to a topic ($\varphi_t(w)$ – represents the probability of word w in topic t). It should be noted that the contributions of each word to the topics does not constitute a distribution, i.e. the sum of contributions does not add up to 1. Assuming the number of topics T, then a simple word-to-word measure is defined by the formula below.

$$LDA-w2w(w,v)=\sum_{t=1}^{T}\varphi_t(w)\varphi_t(v)$$

More global text-to-text similarity measures could be defined in several (Rus, Niraula, & Banjade, 2013).

The last method we present is a semantic similarity method based on **the Quadratic Assignment Problem (QAP).** The QAP method aims at finding an optimal assignment from words in text A to words in text B, based on individual word-to-word similarity, while simultaneously maximizing the match between the syntactic dependencies of the matching words.

The Koopmans-Beckmann (1957) formulation of the QAP problem best fits our purpose. The goal of the original QAP formulation, in the domain of economic activitiy, was to minimize the objective function QAP shown below where matrix F describes the flow between any two facilities, matrix D indicates the distances between locations, and matrix B provides the cost of locating facilities to specific locations. F, D, and B are symmetric and non-negative.

$$\min QAP(F,D,B)=\sum_{i=1}^{n}\sum_{j=1}^{n}f_{i,j}d_{\pi(i)\pi(j)}+\sum_{i=1}^{n}b_{i,\pi(i)}$$

The $f_{i,j}$ term denotes the flow between facilities i and j which are placed at locations $\pi(i)$ and $\pi(i)$, respectively. The distance between these locations is $d_{\pi(i)\pi(i)}$. In our case, F and D describe dependencies between words in one sentence while B captures the word-to-word similarity between words in opposite sentences. Also, we have weighted each term in the above formulation and instead of minimizing the sum we are maximizing it resulting in the formulation below.

$$\max QAP(F,D,B)=\alpha\sum_{i=1}^{n}\sum_{j=1}^{n}f_{i,j}d_{\pi(i)\pi(j)}+(1-\alpha)\sum_{i=1}^{n}b_{i,\pi(i)}$$

A comparison table of the QAP performance on the MSRP corpus, used for comparison purposes as it is the most used paraphrase corpus, is shown in Table 1. QAP offers best results in terms of accuracy (% correct predictions).

Method	**Accuracy**	**F-Score**
All Paraphrase Baseline	66.5%	79.9%
(Corley & Mihalcea, 2006)	71.5%	81.2%
(Qiu, Kan & Chua, 2006)	72.0%	81.6%
(Fernando and Stevenson, 2008)	74.1%	82.4%
(Kozareva and Montoyo, 2006)	76.6%	79.6%
(Socher et al, 2011)	76.8%	83.6%
(Madnani, Tetreault & Chodorow, 2012)	77.4%	**84.1%**
QAP	**77.6%**	83.6%

Table 1. Comparison of the QAP solution with other state-of-the-art methods.

Discussion and Conclusions

Semantic similarity has become a mainstream approach to the challenging problem of natural language understanding as evidenced by the large number of research papers, data sets, and more recently shared task evaluation campaings such as the Semantic Textual Similarity task at SemEval (Agirre et al., 2012). Among the challenges facing the semantic similarity research community, we would mention the lack of a more precise definition of what a paraphrase is. As noted by Rus, Banjade, and Lintean (2014), there is a big discrepancy between the traditional definition of a paraphrase and the loose definition(s) used by the natural language processing community. One of the big discrepancies refers to the level of common words between the two texts being considered. Traditionally, a paraphrase means re-stating in different words. In contrast, paraphrase identification data sets show a surprisingly high level of lexical overlap, which many times is justifiable such as when asking learners to paraphrase science texts (Rus, Banjade, & Lintean, 2014). Another challenge or opportunity for the semantic similarity research community is the need for more consistency in terms of annotation, e.g. the levels of annotation granularity, of data sets.

In summary, we recommend the following improvements when building resources for semantic similarity:

• A crisper definition of paraphrase is necessary, eventually conditioned by context and what real data indicates.

• There is a need to unify at some degree the set of annotation guidelines for an easier comparison of results across them. The unified guidelines should specify the number and type of labels for annotating instances.
• An unified annotation type should be adopted: expert annotation vs. crowdsourcing vs. a mix in which at least a good portion of the data is expert annotated and the rest crowdsourced.
• The exact choice of pre-processing steps could have a big impact on the overall outcome of a more complex approach. Data creators should provide pre-processed versions of the data sets and not only raw text.
• Data creators should provide both "natural" distributions of instance labels as well as balanced versions in which all labels are equally distributed. Some of the existing data sets do this already. Furthermore, data creators should provide data sets or subsets of the original data set that are equally distributed in terms of lexical overlap. That is, the data sets should contain an equal number of instances in which the lexical overlap is say 10%, 20%, and so on up to 90%.
• Ideally, the data set should be created to address as many of the phenomena related to the target task as possible. For instance, pronoun resolution is important for paraphrase identification (Regneri & Wang, 2012) and so at least a certain number of instances should cover this problem and other important issues such as negation, temporal aspects, numerical reasoning, and broader context.

Acknowledgments
This research was supported in part by Insti-tute for Education Sciences under award R305A100875. Any opinions, findings, and conclusions or recommendations expressed in this material are solely the authors' and do not necessarily reflect the views of the sponsoring agency.

References

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

Dagan, I., Glickman, O., and Magnini, B. 2005. The PASCAL recognizing textual entailment challenge. In Proceedings of the PASCAL Workshop.

Dolan, W.B., Quirk, C., and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

Lintean, M., & Rus, V. (2009). Paraphrase Identification Using Weighted Dependencies and Word Semantics. Proceedings of the 22st International Florida Artificial Intelligence Research Society Conference. Sanibel Island, FL.

Lintean, M., Moldovan, C., Rus, V., & McNamara D. (2010). The Role of Local and Global Weighting in Assessing the Semantic Similarity of Texts Using Latent Semantic Analysis. Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference. Daytona Beach, FL.

Lintean, M.C. (2011). Measuring Semantic Similarity: Representations and Methods (Doctoral dissertation). The University of Memphis, Memphis, TN.

Lintean, M., Rus, V., & Azevedo, R. (2011). Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor, International Journal of Artificial Intelligence in Education, 21(3), 169-190.

McCarthy, P.M. and McNamara, D.S. (2008). User-Language Paraphrase Corpus Challenge, online, 2008.

Regneri, M., Wang, R. (2012). Using Discourse Information for Paraphrase Extraction. In: Proceedings of EMNLP-CONLL, pp. 916–927.

Rus, V. & Graesser, A.C. (2006). Deeper Natural Language Processing for Evaluating Student Answers in Intelligent Tutoring Systems, Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06).

Rus, V., Lintean, M. (2012). "A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics", Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June, 2012, Montreal, Canada.

Rus, V., Niraula, N., & Banjade, R. (2013). Similarity Measures based on Latent Dirichlet Allocation. The 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013), March 24-30, Samos, Greece.

Rus, V., Lintean, M., Banjade, R., Niraula, N., & Stefanescu, D. (2013). SEMILAR: The Semantic Similarity Toolkit. The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), August 4-9, Sofia, Bulgaria.

Rus, V., Banjade, R., & Lintean, M. (2014). On Paraphrase Identification Corpora, Proceeding on the International Conference on Language Resources and Evaluation (LREC 2014).

Delphine Bernhard, Iryna Gurevych: 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, June 2008.

Rodney D. Nielsen, Wayne Ward, James H. Martin, and Martha Palmer. (2008). Annotating students' understanding of science concepts. In Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08), Marrakech, Morocco, May 28-30, 2008. Published by the European Language Resources Association, (ELRA), Paris, France.