On Paraphrase Identification Corpora

Vasile Rus, Rajendra Banjade, and Mihai Lintean
Department of Computer Science, Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
E-mail: vrus@memphis.edu, rbanjade@memphis.edu

Abstract

We analyze in this paper a number of data sets proposed over the last decade or so for the task of paraphrase identification. The goal of the analysis is to identify the advantages as well as shortcomings of the previously proposed data sets. Based on the analysis, we then make recommendations about how to improve the process of creating and using such data sets for evaluating in the future approaches to the task of paraphrase identification or the more general task of semantic similarity. The recommendations are meant to improve our understanding of what a paraphrase is, offer a more fair ground for comparing approaches, increase the diversity of actual linguistic phenomena that future data sets will cover, and offer ways to improve our understanding of the contributions of various modules or approaches proposed for solving the task of paraphrase identification or similar tasks. We also developed a data collection tool, called Data Collector, that proactively targets the collection of paraphrase instances covering linguistic phenomena important to paraphrasing.

Keywords: paraphrase identification, paraphrase corpora, semantic similarity

1. Introduction

We analyze in this paper a number of data sets proposed over the last decade or so for the task of paraphrase identification. In particular, we analyze data sets developed since the public release of the first major corpus for paraphrase identification: The Microsoft Research Paraphrase corpus (MSRP; Dolan, Quirk, & Brockett, 2004; Dolan & Crocket, 2005).

We will focus primarily on data sets proposed for the task of paraphrase identification although data sets for related tasks have been proposed, e.g. data sets used for the tasks of recognizing textual entailment (RTE; Dagan, Glickman, & Magnini, 2005; Rus & Graesser, 2006) and elaboration detection (McCarthy & McNamara, 2008). Furthermore, the emphasis of the paper is on sentential paraphrases as they were the primary target of recent major efforts although finer or coarser-grain paraphrases such as phrase level and paragraph level paraphrases have been studied as well although to a lesser extent (Lin & Pantel, 2001; Lintean, Rus, & Azevedo, 2011).

The importance of the paraphrase identification task and of the broader problem of semantic similarity is evidenced by the recent Semantic Textual Similarity shared task that attracted 34 teams (STS; Agirre et al., 2013) and by the release of semantic similarity toolkits such as SEMILAR (Rus et al., 2013).

We argue that the quality of the available corpora for the task of semantic similarity corpus can be significantly improved to serve the important research purposes of identifying the best approaches and understanding the strengths and weaknesses of the various approaches or parts of complex approaches. Current data sets do not facilitate our understanding of the contributions of various components of an approach to the task of paraphrase identification and have limited capacity with respect to fairly compare the overall performance of various approaches. For instance, current data sets only provide raw texts as input whereas we show that depending how these raw texts are preprocessed can make a big difference in the outcome of even simple approaches. Other researchers sporadically expressed concerns about existing paraphrase corpora (Weeds, Weir, & Keller 2005; Zhang & Patrick 2005). However, there is no systematic analysis such as the one presented here across data sets and no solid set of recommendations that can be used as a reference by future developers of paraphrase identification data. The major outcome of our analysis is such a set of recommendations to inform the construction of future data sets.

While we attempt to analyze as many data sets as possible, this analysis is by no means exhaustive for many reasons including space reasons. Furthermore, the paper has a slight emphasis on the MSRP corpus as it is the most well understood data set given that it has been in use for an extended period of time, i.e. almost a decade, that is, much longer than any other data set. We also describe in more detail the User Language Paraphrase corpus (ULPC; McCarthy & McNamara, 2008), which we explored in more detail in the past.

The rest of the paper starts with a discussion of the standard definition of the concept of paraphrase and relate that to the concept of paraphrase defined by Natural Language Processing (NLP) researchers and to the evidence provided by existing paraphrase data sets. Next, we present a number of paraphrase corpora and compare them. We then provide a set of recommendations for improving the process of building and using paraphrase data sets for evaluation purposes in the future. We end the paper with describing the Data Collector, a data collection and annotation tool that proactively targets the collection of instances containing linguistics phenomena important to paraphrasing.

2. What Is A Paraphrase?

A quick search with the query *What is a paraphrase?* on a major search engine reveals many definitions for the concept of paraphrase. Table 1 presents a sample of such definitions. From the table, we notice that the common feature in these definitions is "different/own words." We call this the dictionary definition of paraphrase.

The definition of a paraphrase by NLP researchers varies from a no explicit, very loose definition implied by somehow controversial annotation guidelines (Dolan, Quirk, and Brockett, 2004) to a complex, although still loose, definition of a paraphrase which includes up to 10 dimensions (McCarthy & McNamara, 2008).

Source	***Definition***
Wikipedia	*a restatement of a text or passage using different words.*
WordNet	*express the same message in different words; rewording for the purpose of clarification.*
Purdue's OWL	*your own rendition of essential information and ideas expressed by someone else, presented in a new form.*
Pearson's Glossary	*to record someone else's words in the writer's own words.*

Table 1. Definitions of a paraphrase.

Indeed, the guidelines for the MSRP corpus do specify a loose definition of "semantically equivalence" for paraphrases. A consequence of this somehow loose definition is the emergence of a double standard for judging paraphrases. On one hand, two sentences are considered paraphrases of each other if and only if they are semantically equivalent, i.e. they both convey the same message with no additional information present in one sentence but not the other. We call these precise paraphrases. An example of two sentences that are semantically equivalent is given below (an actual instance from MSRP).

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.*

On the other hand, two sentences are judged as a paraphrase if they convey roughly the same message (minor details being different is acceptable). In this case, the paraphrase relation can be looked at as a bidirectional entailment relation (Text A entails Text B and Text B entails Text A). To exemplify such loose paraphrases, we show below a pair of sentences that has been tagged as a true paraphrase in MSRP:

Text A: *Ricky Clemons' brief, troubled Missouri basketball career is over.*

Text B: *Missouri kicked Ricky Clemons off its team, ending his troubled career there.*

In this example, the first sentence specifies that the career of Mr. Clemons was brief, while the second sentence specifies the reason why Mr. Clemons' career is over. The MSRP corpus contains both types of sentential paraphrases, i.e. precise and loose paraphrases.

Besides the above double standard when it comes to judging paraphrases, we observed another interesting pattern in several existing paraphrase data sets: they tend to have high lexical overlap, i.e. the sentences in a paraphrase instance share many words in common. It should be noted that this pattern of high lexical overlap defies the standard, dictionary definition of a paraphrase which, as we learned earlier, is about conveying the same meaning using "different words." While the standard definition of a paraphrase does not specify the amount of words that must be different, the typical understanding is that most, if not all, of the words must be different.

We argue below that the unexpected high lexical overlap in some of the existing data sets is a consequence of how these data sets were built. Furthermore, it reveals a more fundamental problem with the dictionary definition of a paraphrase, in particular sentence level paraphrases, in the sense that the "different words" requirement in the standard definition may be too strong. In certain contexts, such as learning about science topics, the different words requirements with the understanding of most words being different is too strong.

Indeed, while the dictionary definition of a paraphrase seems to be quite clear with respect to using different words (and preserving the meaning), one particular type of paraphrase, sentence-level paraphrase, does not seem to follow this definition as evidenced by existing data sets, i.e. MSRP and ULPC, built for the purpose of studying sentence-level paraphrases. In MSRP, the average simple word overlap (number of common tokens divided by the average length of the two sentences) equals 68% while in the ULPC corpus the average simple word overlap is 57.65%. If words were lemmatized first, the overlap increases slightly to 69.5 % for MSRP and to 57.65% for the ULPC corpus.

Furthermore, an interesting if not provocative finding is that a simple lexical overlap approach yields much better results that many sophisticated approaches. The S.M.W.B.C.U.N.F. approach (we will explain shortly the meaning of this label) in Table 2 yields an accuracy of 74.32% which is greater than the accuracy of the majority of the 14 approaches, many of which are sophisticated, listed on the ACL wiki's entry for paraphrase identification.

It is important to note that the best simple lexical overlap results have been obtained by optimizing over combinations of pre-processing steps. We considered 1,152 combinations of preprocessing steps resulting in as many variations of the simple lexical overlap approach. To illustrate how we obtained the 1,152 combinations, we label

each instance of the simple lexical overlap approach with an unique id of the form shown below.

MethodName=(O|S).(A|M).(*P*|*W*|*C*|*S*).(*W*|*B*|*P*).(*S*|*I*).(*U*|*B*).(*I*|*E*|*N*).(*F*|*N*)

Method	**Accuracy**	**Precision**	**Recall**
O.A.P.B.C.U.N.N.	.7258	.7705	.8370
O.A.P.B.I.U.N.N.	.7403	.7538	.9050
S.M.W.B.C.U.N.F.	**.7432**	.7600	.8971
O.A.P.B.I.B.I.N.	*.6783*	*.6947*	*.9207*
S.A.S.B.I.U.N.N.	**.6433**	**.6156**	**.9267**
O.A.P.B.I.B.N.F.	*.6072*	*.6066*	*.8022*

Table 2. Results for a simple lexical overlap method with various combinations of pre-processing steps for the MSRP test (top four rows) and ULPC test (bottom two rows). We highlight in bold/italics the combination of preprocessing steps leading to best/worst accuracy, respectively.

The first letter in a method's name indicates whether we used OpenNLP package (O) or Stanford NLP package (S). The second letter indicates the type of normalization when computing the lexical overlap: average length (A) versus maximum length (M). The remaining pre-processing steps indicate: (1) the tokens used from the original sentences (P means we compared all tokens, including punctuation; W means we excluded punctuation; C means content words only; S means all words, excluding the stop words), (2) what form of the retained tokens was used (W - original raw form, B - base form, P means we compared only words that have the same part-of-speech and same base form), (3) case sensitivity (S) or insensitivity (I), (4) unigrams (U) or bigrams (B), (5) type of global weight used for each token (I means IDF, E means entropy-based, or N means weight of 1), and (6) type of local weight used (F means word type frequency, N means local weighting of 1).

It is important to add that Table 2 shows extreme results (best and worst) obtained with various versions of the lexical overlap method. The wide variation in these results suggests that data creators should provide standardized pre-processed versions of the data and ask users of the data set to report results on such pre-processed data as well besides standard evaluations on the raw data. The very competitive results obtained with such a simple method is at some extent a consequence of the high lexical overlap characteristics of the data and of the skewness of the data set towards positive instances. It should be noted that the MSRP instance distribution (dominated by positive instances) is in contrast with recent data sets (see Table 3) such as Student Response Assessment (SRA; Dzikovska et al., 2013) and Rekneri and Wang (2012) which contain more negative instances.

Another interesting effect of the high lexical overlap is the fact that modifiers seem to weight more in deciding whether two sentences are paraphrases, which is counterintuitive as the main content words, not modifiers, should weight more (Lintean, 2011). This is yet another consequence of the high lexical overlap pattern.

While the high lexical overlap of the paraphrases in the MSRP corpus can be explained by the protocol used to create the corpus - same keywords were used to retrieve same stories from different news sources on the web, one could further argue that avoiding the high word overlap issue in sentential paraphrasing would be hard in the news domains where high concentration of named entities is often the case. For instance, given an isolated sentence it would be quite challenging to omit/replace some core concepts when trying to paraphrase. Here is an example of a sentence (instance 735 in the MSRP corpus), *Counties with population declines will be Vermillion, Posey and Madison.*, which would be hard to paraphrase using many other/different words. The difficulty is due to the large number of named entities in the sentence. Indeed, replacing these named entities with new words is hard if not impossible without changing the meaning of the sentence substantially. The paraphrase of the above example in the MSRP corpus is *Vermillion, Posey and Madison County populations will decline.*

The same pattern of high lexical overlap is present in yet another corpus, the ULPC corpus, which contains real student paraphrases of biology textbook sentences collected from experiments with iSTART, an intelligent tutoring system that teaches students reading strategies (McNamara, Boonthum, et al., 2007). The high overlap between a student paraphrase sentence and the original textbook sentence in the ULPC corpus is evidence that when middle-school or high-school students learning biology are asked to paraphrase biology sentences they reuse more than half of the words in the reference sentence.

This is yet another argument that requiring "different words" in a sentence-level paraphrase is too strong of a requirement in certain domains such as paraphrasing science texts or even news texts and that high lexical overlap should be acceptable contrary to the standard, dictionary definition of paraphrase.

There are other contexts in which high lexical overlap is acceptable. For instance, in the context of a conversation rephrasing the speaker's most recent turn by simply repeating it (high lexical overlap) may be acceptable as a form of double-checking the understanding of the speaker's message. In contrast, in some contexts, such as essay writing or writing in general where the exact choice of words, i.e. exact form of expression, is key, plagiarism is a concern and therefore high lexical overlap without acknowledging the source may not be acceptable.

It is beyond the scope of this article to provide a final answer with respect to whether high lexical overlap should be acceptable or not in sentential paraphrases. We are simply raising the issue for further community discussion. A clear definition of a paraphrase is indeed hard to find as Barzilay's survey (2003) indicates. However, one hopes that a decade later and dozens of studies and datasets later some progress towards a crisper definition of a paraphrase has been made.

Corpus	Size	Distribution	Number of labels
MSRP	5,801	Total: 3900(67%) –paraphrase Training: 2753 (67.54%) – paraphrase. Test: 1147 (66.5%) – paraphrase.	2 (1 – paraphrase, 0 – nonparaphrase) Expert Annotation
ULPC	1,998	Training: 1,012 (50.7%) Validation: (337 items, 16.9% Test - (649 items, 32.5%)	6 Ratings (1-3 – no paraphrase, 1 having higher confidence; 4-6 – paraphrase, 6 having highest confidence) Expert Annotation
QP (Bernhard & Gurevych, 2008)	7,434	7,434 true paraphrases for 1,000 target questions (7.434 paraphrased questions per target questions)	Wiki-based, community-based annotation
SEMILAR (Rus et al., 2012)	700	344 (49%) – Paraphrase based on the MSRP annotation. 442 (63%) – True paraphrase based on annotation done by SEMILAR annotators.	2 (0 – non-Paraphrase, 1- Paraphrase) Expert Annotation
SRA(Dzikovska et al., 2013)	14,228	Train: correct – 1898 (0.27); incorrect – 5237 (0.73) Test: correct – 2971 (0.42); incorrect – 4122 (.58)	2-way (they also provide a 3way categorization) Heuristic annotation
STS (Agirre et al., 2013)	2,250	750 (news), 189 (Framenet-Wordnet glosses), 561 (OntoNotes-Wordnet glosses), 750 (MT evaluation). [0-1]: 453 (20.133%) (1- 2]: 249 (11.067%) (2- 3]: 247 (10.978%) (3- 4]: 572 (25.422%) (4- 5]: 729 (32.400%)	6 (5 – identical 4 – Strongly related 3 – Related 2 – Somewhat Related 1 - Unrelated 0 – Completely unrelated) Crowdsourced annotation
RW (Rekneri & Wang, 2012)	1,992	590 true paraphrases [the sum of their 158 paraphrases, 238 containment cases, 194 related cases] and 1402 unrelated	2-way Expert Annotation

Table 3. Summary of existing paraphrase corpora.

3. Paraphrase Corpora

One of the most important legacies of the MSRP corpus is the inspiration it generated for other researchers to study and develop data sets for paraphrase research in particular and for other semantic similarity tasks. In fact, some of the corpora developed afterwards (re-)use MSRP as a source, e.g. the STS pilot challenge in 2012 includes a portion of the MSRP corpus. Table 3 summarizes the major existing paraphrase data sets. As it can be noticed, the various data sets vary in their annotation, size, instance distribution, sources, and type of annotation. We discuss briefly each of these data sets next.

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) and a test set (1,725 pairs). The number of average words per sentence (sentence length) 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.

The User Language Paraphrase Corpus (ULPC; (McCarthy and McNamara 2008)) contains pairs of target-sentence/student response texts. The student responses were collected from experiments with the intelligent tutoring system iSTART. Students were shown individual sentences collected from biology textbooks and asked to paraphrase them. These pairs have been evaluated by expert human raters along 10 dimensions of paraphrase characteristics. The ”Paraphrase Quality bin” 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 1,998 pairs, 1,436 (71%) were classified by experts as being paraphrases. A quarter of the corpus is set aside as test data. The average words per sentence is 15.

The Question Paraphrase corpus (Bernhard & Gurevych, 2008) 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. 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 1,000 target questions for which there are on average 7.434 paraphrased questions. There are no explicit false paraphrase instances.

The SEMILAR corpus (formerly known as SIMILAR; 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 is 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 answers related to specific tutorial questions from different science domains. There are 56 questions and 3,000 student answers from the so-called BEETLE corpus, 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.

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 also used string similarity for selection of instances. We only describe here the STS CORE corpus as its input is pure text. 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 2,000 sentence pairs from previously existing paraphrase datasets and machine translation evaluation resources. The test data also comprised 2,000 sentences pairs from 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 was rated on a 0-5 scale (low to high similarity) by human judges recruited through Amazon Mechanical Turk.

Regneri and Wang (2012) built a dataset (which we will label RW) starting with 2,000 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.

Other related data sets are the one in Rus and Graesser (2006) and Cohn, Callison-Burch, and Lapata (2008).

Other data sets for paraphrase exist but they do not fit in the general category of sentence-level paraphrases, the focus of our analysis. For instance, Potthast and colleagues (2010) created that PAN corpus which contains paragraph-size texts and Lintean, Rus, and Azevedo (2011) describe another paragraph-level paraphrase corpus.

As can be seen, there is a myriad of data sets with a large variety of distributions, annotation styles, data sources, etc. This diversity is valuable but at the same time it makes it difficult to understand the benefits of approaches using one data set or another. Also, it is hard to fairly compare approaches when there is so much variation in the way the data sets are created. Therefore, it is imperative that the process of building paraphrase datasets be standardized while keeping diversity that is useful. For instance, the definition of a paraphrase must be more precise and the annotation guidelines must somehow converge while the source of data can be diverse. The next section presents a set of recommendations towards improving the process of creating data sets for paraphrase identification of other semantic similarity tasks.

4. Recommendations

Our investigation has helped us articulate a number of recommendations for future data collection and annotation efforts, which are meant to improve the quality of the data sets in ways to help answer critical research questions and to fairly compare approaches or assess the impact of particular components in a more complex approach.

The set of recommendations is provided below.

- 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 set developers should provide pre-processed versions of the data sets and not only raw text. Alternatively, researchers may use the raw text version in which case they must report precisely the pre-processing steps. However, without them releasing their pre-processed data small variations might still exist if someone tries to replicate their description of pre-processing steps. Therefore, the suggestion of having data set developers provide pre-processed versions is preferred.
- Data set developers 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%.
- Ratings should be finer grain, not binary, so that researchers can assess their methods at a finer level.
- Data sets should cover a broad range of linguistic phenomena that are known to be important to paraphrase detection.
- 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.

5. Data Collector and Annotation Tool

As a way to mitigate the tendency of text similarity corpora, many of which are collected automatically such as MSRP, to be unbalanced and biased towards limited linguistic phenomena we developed a data collection and annotation tool that proactively targets a wide range of linguistic phenomena important to paraphrasing. Indeed, because existing data sets are unbalanced and cover a limited range of linguistic phenomena, methods developed and tuned using existing data sets work well for the evaluation corpus but perform poorly in practical applications. From our experience, we learned that student answers collected from experiments with the intelligent tutoring system Deeptutor (www.deeptutor.org) are sometimes very difficult to automatically evaluate against expert's answers because some linguistic phenomena were not appropriately covered in semantic similarity data sets used for training our methods. For example, about 5% of students' utterances contain some form of negation but if we look at the MSRP corpus, negation doesn't seem to have much significance.

Moreover, the inter-rater agreement in the case of MSRP corpus was 84%. Dolan and Brockett (2005) observed that creating a strict guideline for the annotation even dropped the inter-rater agreement but common sense worked well. So, creating and annotating sentence pairs with the help of experts and crowds would lead to a balanced corpus covering all the important linguistic phenomena for paraphrasing. To this end, a web based tool to facilitate the collection and annotation was created: the DataCollector.

The DataCollector tool (http://deeptutor2.memphis.edu/DataCollector/) is an online facility developed with an aim to create balance and linguistically richer paraphrase corpus. Currently user can add new sentence pairs, or paraphrase an existing sentence, and rate the similarity of the sentences in the pair from 0 (completely different meaning) to 10 (exactly the same meaning). The tool is meant to create variability in the collected and annotated dataset and therefore users are guided in their process accordingly although they freedom to focus on whatever linguist phenomena they prefer, with the tool subsequently assuring diversity. For example, an user can add sentence pairs which look similar to each other but they have different meaning and can't be treated as paraphrase. On the other hand, the pairs can have very little or no lexical overlap but they may mean the same or almost the same thing. Users are encouraged to create sentence pairs which are diverse in features. They are also asked to select which is the most prominent or deciding feature in the paraphrase they added to the data set.

Some of the cases the paraphrase identification systems are expected to handle are (but not limited to) shown in Table 4.

To conclude, we hope that our set of recommendations will be used as a reference point for the development of future data sets for semantic similarity tasks. The ultimate goal is to develop data sets that further our understanding of the phenomena and proposed solutions.

Acknowledgements

This research was supported in part by IES Award# R305100875A. 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

Agirre E., Cer D., Diab M., Gonzalez-Agirre A., Guo W. (2013). *SEM 2013 shared task: Semantic Textual Similarity The Second Joint Conference on Lexical and Computational Semantics.

Barzilay, R. (2003). *Information Fusion for MultiDocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University, New York, NY.

Bernhard, D. & Gurevych, I. (2008). 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.

Cohn, T., Callison-Burch, C., and Lapata M. 2008. Constructing Corpora for 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.

Brockett, C. & Dolan, W.B. (2005). Support Vector Machines for Paraphrase Identification and Corpus Construction, in *Third International Workshop on Paraphrasing (IWP2005)*, Asia Federation of Natural Language Processing, 2005.

Dzikovska, M.O., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I. and Dang, H.T. (2013) "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. 13-14 June.

Lin, D. & Pantel, P. 2001. Discovery of Inference Rules for Question Answering. *Natural Language Engineering* 7(4):343-360.

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.

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

McNamara, D.S.; Boonthum, C.; Levinstein, I. B.; and Millis, K. 2007. *Handbook of Latent Semantic Analysis. Erlbaum*, Mahwah, NJ. chapter Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms, 227–241.

Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. Overview of the 1st International Competition on Plagiarism Detection. In Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors, SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), pages 1-9, September 2009. CEUR-WS.org. ISSN 1613-0073.

Regneri, M., Wang, R. (2012). Using Discourse Information for Paraphrase Extraction. In: Proceedings of EMN LP-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., Moldovan, C., Baggett, W., Niraula, N., Morgan, B. (2012). The SIMILAR 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.

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

Weeds, J., Weir, D., & Keller, W. (2005). The distributional similarity of sub-parses. *In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 7–12, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Zhang, Y. & Patrick, J. (2005). Paraphrase identification by text canonicalization. In *Proceedings of the Australasian Language Technology Workshop*.

Text features	Example
High lexical overlap but not paraphrase	A: "There's nothing we can do to stop" the water flow, Stegman said. B: "Right now, there is nothing we can do," Stegman said. (MSRP-Train #3171)
Low lexical overlap but paraphrase	A: The shares of the company dropped. B: The organization's stock slumped.
Varying length	A: "University will host a conference next year", said John Smith. B: John Smith, vice president of academic affairs, said that the university will host a conference sometime next year.
Negation	A: Cory doesn't like tomato juice. B: Cory likes tomato juice.
Meaning in context	A: He warned that his party can boycott the election. B: He said that his party can boycott the election.
Extra information	A: The rainy season is good for farmers but it's annoying to some people. B: The rainy season is good for farmers.
Temporal information	A: The internet connection was dead from morning to evening. B: There was no internet connection for 12 hours.
Numerical data	A: The current price of share doubled in last five days which was $10 last week. B: The current price of share is $20.
Requiring world knowledge	A: Barack Obama recently visited South Africa. B: The president visited South Africa.
Speech act	A: Don't dare to move ahead! B: Don't step forward.
Anaphora	A: She announced the merger deal. B: Kristina announced the merger deal.
Named entity	A: The head of the nations attended UN general convention in New York. B: The head of the nations attended United Nations general convention in the USA.
Phrasal verbs	A: He showed up late. B: He arrived late.
Syntactic features	A: I watched a documentary before I went to bed. B: Before I went to bed, I watched a movie.
Comparative	A: Summer is better than winter. B: Winter is better than summer
Quantifiers	A: A large number of people are waiting in the queue. B: There is a long queue.
Adjectives	A: She is wearing a white t-shirt. B: She is wearing a blue t-shirt.
Modality	A: John will possibly go to Atlanta. B: John will go to Atlanta.
Metaphoric	A: Time is a thief. B: Time passes quickly.

Table 4. Linguistic phenomena targeted by the Data Collector tool