The SIMILAR Corpus: A Resource To Foster The Qualitative Understanding of Semantic Similarity of Texts

Vasile Rus, Mihai Lintean, Cristian Moldovan, William Baggett, Nobal Niraula, and Brent Morgan

Department of Computer Science, Department of Psychology, Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
E-mail: vrus@memphis.edu, mclinten@memphis.edu, cmldovan@memphis.edu, nbnraula@memphis.edu, writebill@gmail.com, brent.morgan@gmail.com

Abstract

We describe in this paper the SIMILAR corpus which was developed to foster a deeper and qualitative understanding of word-to-word semantic similarity metrics and their role on the more general problem of text-to-text semantic similarity. The SIMILAR corpus fills a gap in existing resources that are meant to support the development of text-to-text similarity methods based on word-level similarities. The existing resources, such as data sets annotated with paraphrase information between two sentences, do not provide word-to-word semantic similarity annotations and quality judgments at word-level. We annotated 700 pairs of sentences from the Microsoft Research Paraphrase corpus with word-to-word semantic similarity information using both a greedy and optimal protocol. We proposed a set of qualitative word-to-word semantic similarity relations which were then used to annotate the corpus. We also present a detailed analysis of various quantitative word-to-word semantic similarity metrics and how they relate to our qualitative relations. A software tool has been developed to facilitate the annotation of texts using the proposed protocol.

Keywords: word-to-word semantic similarity, paraphrase identification, entailment recognition

1. Introduction

We describe in this paper our effort to fill a gap in existing resources for the study of semantic similarity of texts. We have designed a protocol and created an annotated data set to foster a deeper and qualitative understanding of word-to-word semantic similarity measures together with their role on the more general task of assessing the semantic similarity of texts (containing more than one word). An example of a text-to-text semantic similarity task is the task of paraphrase identification (Dolan, Quirk, and Brockett, 2004).

The semantic similarity approach, as a practical alternative to the full understanding approach to the task of natural language understanding (Rus & Lintean, submitted), has been successfully applied to a series of fundamental text-to-text similarity tasks in natural language processing: paraphrase identification (Dolan, Quirk, and Brockett, 2004), recognizing textual entailment (Dagan, Glickman, & Magnini, 2005; Rus & Graesser, 2006), and elaboration detection (McCarthy & McNamara, 2008). These fundamental tasks are in turn important to a myriad of real world applications such as providing evidence for the correctness of answers in Question Answering (Ibrahim, Katz, & Lin, 2003), increase diversity of generated text in Natural Language Generation (Iordanskaja, R. Kittredge, & A. Polgere, 1991), assessing the correctness of student responses in Intelligent Tutoring Systems (Graesser, Hu, McNamara, 2005), and identifying duplicate bug reports in Software Testing (Rus et al., 2009). Table 1 provides examples of text pairs from semantic similarity tasks proposed by various research groups over the last decade.

Much research has been dedicated to proposing word-to-word similarity metrics (Pedersen, Patwardhan, and Michelizzi, 2004) and more recently to developing methods to compute the semantic similarity of larger texts. Among the latter, a particular set of methods that address the larger text-to-text similarity problem are those that rely on word-level similarity metrics (e.g. the similarity of two sentences or paragraphs; Corley & Mihalcea, 2005; Lintean et al., 2010) and which we call compositional methods as they are based on the principle of compositionality. The compositional principle states that the meaning of longer texts can be composed from the meaning of its parts, i.e. words.

To the best of our knowledge existing methods to solve the text-to-text similarity problem using word-level similarities limit themselves to a quantitative analysis of the overall method's performance on a given text-to-text similarity task, e.g. paraphrase identification, as opposed to a more detailed quantitative and qualitative understanding of the word-to-word similarity metrics and their impact on the text-to-text similarity method proposed. How does the average similarity score between words that are deemed similar beyond any doubt compare to the average similarity score between words that are deemed similar in some context? For instance, what is the qualitative difference between a similarity score of 0.90 and a score of 0.70 (we assume normalized similarity scores only)? What about between a score of 0.45 and a score of 0.55? Also, it is not known at what extent these word-level metrics capture more than lexical information, e.g. context and world knowledge. We take a first step towards a better understanding of word-to-word similarity metrics and their actual impact on methods using these metrics.

To this end, we propose a protocol that maps existing

ID	Text 1 **(assumed to be True for tutoring and RTE data)**	Text 2	Source/Relation
1	**Expert Answer:** The force of the earth's gravity, being vertically down, has no effect on the object's horizontal velocity	***Student Input:*** *The horizontal component of motion is not affected by vertical forces*	AutoTutor/True Paraphrase
2	**Textbook Sentence**: A glacier's own weight plays a critical role in the movement of the glacier.	**Student Input**: A glacier's movement depends on its weight.	iSTART/True Paraphrase
3	The procedure is generally performed in the second or third trimester.	*The technique is used during the second and, occasionally, third trimester of pregnancy.*	MSR/True Paraphrase
4	**Text**: Deployment of Filipino workers in Iraq suspended by Philippine president due to repeated kidnappings.	**Hypothesis**: *Filippino workers have been kidnapped by the Philippine president.*	RTE/False Entailment

Table 1. *Examples of text pairs from four different datasets: AutoTutor, iSTART, Microfost Research Paraphrase (MSR) corpus, and Recognizing Textual Entailment (RTE) corpus.*

word-to-word similarity metrics onto qualitative judgments of similarity such as CLOSE (the words are similar beyond any doubt, e.g. *student* and *learner*), RELATED (the words are related but they are not quite similar, e.g. *boxing* and *fight*), CONTEXT (the words are matched within the context of the texts to be assessed, e.g. *totalling* and *volume* – see the whole context later), and KNOWLEDGE (world or domain knowledge is needed to match the words, e.g. *retailer* and *WalMart*). These qualitative judgments are then related to existing quantitative word-to-word similarity metrics for a better understanding and interpretation of the metrics.

The protocol was designed in the context of qualitative assessments of the similarity of two texts. That is, judges were shown two texts which might or might not be semantically similar, e.g. paraphrases, and asked to match words and indicate the reason such as CLOSE, RELATED, CONTEXT, KNOWLEDGE. A default NONE value is assigned to unmatched words. Identical words (in their raw form) in both sentences were deemed perfectly similar and annotated automatically with the label IDENTICAL.

We chose as our starting data set the Microsoft Research Paraphrase corpus (MSRP; Dolan, Quirk, and Brockett, 2004) used to evaluate methods addressing the task of paraphrase identification. The corpus has been widely used by many research groups (Corley & Mihalcea, 2005; Lintean & Rus, 2009; Lintean et al., 2010) and therefore would allow us to compare the results of word matching by human annotators with the machings proposed by the automated methods. We have asked the human experts to pair words greedily as well as optimally. The greedy annotation was necessary in order to emulate existing automated greedy methods (Corley and Mihalcea, 2005; Lintean et al., 2010) which would allow for a direct comparison with human greedy judgments. In the greedy annotation, we asked humans to consider one word at a time in one text, say T1, and greedily match it to a word in the other text, T2, without considering the whole text T1 as a context. Optimal annotation of similar words was based on human judges' full understanding of the texts.

We annotated as of this writing 700 pairs of sentences from the MSRP corpus which consists of 29,771 tokens (words and punctuation) of which 26,120 are true words and 17,601 content words. The 700-pair dataset also contains 12,560 true relations (a true relation is of any type except NONE) identified when greedily identifying similarities from T1 to T2 (target words were selected from T1) and 12,345 true relations identified when greedily annotating from T2 to T1. For the optimum annotation, 15,692 relations were identified. We report a detailed analysis of the so obtained corpus, called the SIMILAR corpus, and compare the human annotations with results obtained by matching words using the word-to-word semantic similarity measures in the WordNet Similarity library (Pedersen, Patwardhan, and Michelizzi, 2004) as well as using Latent Semantic Analysis (LSA; Landauer et al., 2007).

A semantic annotation tool was also developed that allowed our experts to easily annotate the SIMILAR corpus. The tool offers an user-friendly interface which tremendously speeds up the transfer of the proposed annotation protocol to new texts, in any language, and also offers great productivity advantage allowing for annotating more text per unit of time. If the paper is accepted, both the corpus and the annotated data set will be available at our website: www.semanticsimilarity.org.

The rest of the paper is organized as in the following. The next section presents related work on semantic similarity with an emphasis on compositional approaches based on word-to-word similarity metrics. Section 3 describes in details the guidelines for greedy annotation while section 4 presents guidelines for optimum annotation. The annotation tool is briefly described in section 5. The details of the SIMILAR corpus are presented in the following section. The Conclusions section ends the paper.

2. Related Work

Assessing the semantic similarity of texts has been explored at different levels of granularity: word-to-word, sentence-to-sentence, paragraph-to-paragraph (Rus, Lintean, & Azevedo, 2009), and document-to-document (see Information Retrieval work; Salton, Wong, & Yang, 1975). We focus next on word-to-word similarity and sentence-to-sentence similarity work as it is most relevant to ours.

Word-to-word similarity research culminated with the release of the WordNet similarity package by Pedersen, Patwardhan, and Michelizzi (2004). Other notable work that allows quantifying how similar words are is the Latent Semantic Analysis framework (described below) and more recently Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003). Other frameworks exists which we do not mention due to space limitations.

Extending word-to-word similarity measures to sentence level and beyond has drawn increasing interest in the last decade or so in the Natural Language Processing community. The interest has been driven primarily by the creation of standardized data sets and corresponding shared task evaluation campaigns (STECs) for the major text-to-text qualitative semantic relations of entailment (RTE; Recognizing Textual Entailment corpus by Dagan, Glickman, & Magnini, (2005), paraphrase (MSRP; Microsoft Research Paraphrase corpus by Dolan, Quirk, and Brockett, 2004), and elaboration (ULPC; User Language Paraphrase Challenge by McCarthy & McNamara, 2008).

Assessing the semantic similarity of two texts, T1 and T2, using a compositional approach based on word-to-word semantic similarity metrics has been primarily approached using greedy methods (Corley & Mihalcea, 2005; Lintean & Rus, 2009; Lintean et al., 2010) and more recently an optimal method (Rus & Lintean, in press). We briefly describe these approaches as they are relevant to our corpus annotation effort.

Corley and Mihalcea (2005) presented one of the earliest methods to compute the similarity of two sentences using word-to-word similarity methods. In their method, they computed the similarity of two texts by greedily summing up the maximum similarity of each word in one sentence to any word in the opposite sentence. The individual word-to-word similarities were computed using measures from the WordNet similarity package (Pedersen, Patwardhan, & Michelizzi, 2004) as well as a simple vector space model. They report results on the MSRP corpus. Other notable work is by Rus and colleagues (2008) who addressed the task of paraphrase identification using the MSRP corpus by computing the degree of subsumption at lexical and syntactic level between two sentences in a greedy manner as well.

Assessing the correctness of student contributions in dialogue-based tutoring systems has been approached either as a paraphrase identification task (Graesser, Hu, McNamara, 2005; Graesser, Olney, et al., 2005), i.e. the task was to assess how similar student contributions were to expert-generated answers, or as an entailment task (Rus & Graesser, 2006), in which case the task was to assess whether student contributions were entailed by expert-generated answers. The expert answers were assumed to be true. If a correct expert answer entailed a student contribution then the contribution was deemed to be true as well.

Latent Semantic Analysis (LSA; Landauer et al., 2007) has been used to evaluate student contributions during the dialog between the student and a dialogue-based tutoring system (Graesser, Hu, & McNamara, 2005; VanLehn et al., 2007). In LSA the meaning of a word is represented by a reduced-dimensionality vector derived by applying an algebraic method, called Singular Value Decomposition (SVD), to a term-by-document matrix built from a large collection of documents. A typical dimensionality of an LSA vector is 300-500 dimensions. To compute the similarity of two words the cosine of the words' corresponding LSA vectors is computed (cosine is the normalized dot-product). A typical extension of LSA-based word similarity to computing the similarity of two sentences (or even larger texts) is to use vector algebra to generate a single vector for each of the sentences/texts (by adding up the LSA vectors of the individual words) and then compute the cosine between the resulting sentence/text vectors. Another approach proposed, greedily selects for each word its best match using the cosine of the words' LSA vectors, and then sums the individual word-to-word similarities in order to compute the overall similarity score for the two sentences (Lintean et al., 2010). Our work is mostly relevant to LSA-based approaches using only the latter method as it is the only approach that fits with a compositional model based on word-to-word similarity.

We describe the greedy and optimal methods in more details next. It is important to describe them as our manual annotation tries to emulate them (although the optimal manual annotation is slightly different compared to the optimal automated method).

Greedy Method

In the greedy method, each word in text T1 is paired with every word in text T2 and word-to-word similarity scores are computed according to some metric. For each word in T1, its best matching word in T2 is greedily retained. These greedily-obtained scores are added up using a simple or weighted sum which can be normalized in different ways, e.g. by dividing to the longest text or to the average length of the two texts. The formula we show here is given in equation 1 (from Lintean & Rus, 2009). As one would notice, this formula is asymmetric, i.e. score(T1,T2) ≠ score(T2,T1). The average of the two scores provides a symmetric similarity score, more suitable for a paraphrase task, as shown in Equation 2. Given that identical words occurring in the two texts are perfectly matched, the greedy method matches identical words first.

$$score(T1,T2) = \frac{\sum_{v \in T1} weight(v) * \max_{w \in T2} word - sim(v,w)}{\sum_{v \in T1} weight(v)}$$

Equation 1. *Asymmetric semantic similarity score between texts T1 and T2.*

$$simScore(T1,T2) = \frac{score(T1,T2) + score(T2,T1)}{2}$$

Equation 2. *Symmetric semantic similarity score between*

ID	SENTENCE	TARGET	SEMANTIC RELATION
1	In Nigeria alone, the report estimated that between 100,000 and 1 million girls and women are suffering from the condition.	running	NONE
2	The charges allege that he was part of the conspiracy to kill and kidnap **persons** in a foreign country.	individual	WORD
3	Hearing was partially **restored** by an electronic ear implant.	regained	WORD
4	In Nigeria alone, the report said, as many as 1 million women may be **living with the condition**.	suffering	PHRASE
5	Jeter, who **dislocated his left shoulder** in a collision March 31, took batting practice on the field for the first time Monday.	injury	PHRASE
6	NASA satellite images show that Arctic ice has been shrinking at the rate of nearly 10 percent a decade.	disappearing	CONTEXT
7	Duke and North Carolina have been resolute in their positions against expansion.	oppose	CONTEXT
8	The retailer said it came to the decision after hearing the opinions of customers and associates.	Wal-Mart	WORLD KNOWLEDGE
9	Duke and North Carolina have been resolute in their positions against expansion.	school	WORLD KNOWLEDGE

texts T1 and T2.

Table 2. *Examples of target words (third column), opposite sentences (column two), and qualitative similarity relations (last column).*

The obvious drawback of the greedy method is that it does not aim for a global maximum similarity score. The optimal method (Rus & Lintean, in press) which is described next solves this issue.

Optimal Method

The optimal matching solution (Rus & Lintean, in press) was inspired by the optimal assignment problem which is one of the fundamental combinatorial optimization problems and 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 $w(xy)$, find a matching M from X to Y with maximum weight.

A famous instance of the optimal assignment problem is job assignment which is about assigning a group of workers, e.g. sailors, to a set of jobs (on ships) based on the expertise level, measured by $w(xy)$, of each worker at each job (Dasgupta et al., 2009). By adding dummy workers or jobs we may assume that X and Y have the same size, n, and can be 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 workers and jobs are words from the two sentences to be compared and the weight $w(xy)$ is the word-to-word similarity between words x and y in the two sentences, respectively.

The assignment problem can thus be formulated as finding a permutation π of {1, 2, 3, … , n} for which $\sum_{i=1}^{n} w(x_i y_{\pi(i)})$ is maximum (Dawes, 2011). 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. It is beyond the scope of this paper to present the details of the algorithm.

The method guarantees optimal overall best match. That is, Rus and Lintean (in press) showed how using the Kuhn-Munkres algorithm words in text T1 (the sailors) cab be optimally matched to words in text T2 (the ships) based on how well the words in T1 (the sailors) fit the words in T2 (the ships). The fitness between the words is nothing else but their word-to-word similarity according to some metric of word similarity.

Based on these two categories of compositional semantic similarity approaches that rely on word-to-word similarity metrics, greedy and optimal, we have designed two annotation protocols: greedy and optimal annotation.

3. Greedy Word-to-Word Annotation

As already mentioned, the greedy matching strategy was inspired from automated greedy methods proposed for the task of semantic similarity of short texts. The greedy methods pair a target word in one sentence with all the words in the other sentence and retain the matching word with the highest word-to-word similarity score to the target word regardless of how other words match each other.

If human judges were to emulate this process they would have to consider one individual word from one sentence, called the target word, and try to find a best matching word in the other sentence regardless of how other words would match. This isolation assumption is needed to emulate the word-to-word similarity measures as closely as possible and allow a direct comparison between human judgments and automated methods. Table 2 illustrates how the greedy annotation occurred. It also provides examples for each type of qualitative word-to-word relations we defined. The third columns shows target words, from text T1, and the second column all candidates words from text T2. The other words in T1 are irrelevant in greedy matching. Note the greedy matching needs to be performed in two phases. Phase one means selecting target words from T1 and find best matches in T2. Phase two involves selecting target words from T2 and find best matches in T1.

3.1 The Qualitative Word-to-Word Relations

When selecting the best matching individual word in the

opposite sentence for a given target word, judges must decide whether a matching word exist (or not). If a matching word exists, a judgment on the type of matching needs to be made. A matching word could be a word which is semantically close, based on judge's view, to the target word. Semantically close words are words that are synonyms such as *person* and *individual*, or deemed semantically close beyond any reasonable doubt by a human judge. If words have multiple senses, at least two senses of the two words are semantically close beyond any reasonable doubt). For instance, the words *research* and *study* are semantically close when considering their meaning of *investigating* a particular issue.

In case a semantically close word is not found, a word that is somehow semantically related should be chosen, e.g. *boxing* and *fighting* are semantically related but not semantically close.

These two types of annotations would be sufficient to directly evaluate greedy automated methods against the human greedy judgments. However, we wanted to go beyond that. We decided to include in the annotation protocol several additional types of qualitative semantic relations.

If a target word is not similarly close or related as defined above to any individual word in the other sentence (when considering these words in isolation), it might be the case that the two words could be deemed similar if the context of the matching word (but not of the target word) could help in relating semantically the words. For instance, the target word *totalling* is contextually related to *volume* in the second sentence below if considering the full *context* of the second sentence.

T1: Singapore is already the United States' 12th-largest trading partner, with two-way trade totaling more than $34 billion.

T2: Although a small city-state, Singapore is the 12th-largest trading partner of the United States, with trade volume of $33.4 billion last year.

For the context relation it might be the case that a particular target word cannot be matched against one individual word in the other sentence. It is rather the case that the other sentence entirely implies or suggests the target word in which case the target word is related to the context of entire sentence instead of one particular word. This might be the case also for the next type of relation, KNOWLEDGE.

Sometimes even context is not enough to relate a target word to any other word in the opposite sentence. Word knowledge could help. In the above example, when matching the target word/collocation *city-state* world knowledge is needed to relate it to *Singapore* in the first sentence.

Sometimes a target word, e.g. the collocation *credit_card* in the second sentence below, cannot be matched in any way to a word in the other sentence. In this case, the NONE relation is chosen for the target word.

T1: He said it was a mistake, and he reimbursed the party nearly $2,000.

T2: The governor said the use of the credit card was a mistake, and has since reimbursed the party for the expense.

3.2 Additional Guidelines

Collocations such as *give_up* or *joint_venture* were considered individual words because word-to-word similarity metrics consider them so and therefore similarity scores can be computed between collocations or between a collocation and a simple word.

Numbers were deemed as either semantically close, when identical, or semantically related when representing different values, e.g. 123 and 345 are related.

Temporal markers, such as *today* or *yesterday*, were deemed close, when identical, and related when different.

Pronouns should were deemed close, when identical, and contextually related to a referent when could be linked to the referent in the opposite sentence (or NONE if no reasonable referent was found).

Punctuation had to be matched to an identical punctuation mark in the opposite sentence.

Verbs were matched using their base forms and ignoring inflections. For instance, *go*, *went*, *gone* were all matched with each other.

Auxiliaries, e.g. *has* in *has gone*, were labelled with NONE if the main verb (i.e. *gone*) had no match in the opposite sentence. When the main verb does have a match, the auxiliary was matched with a matching/corresponding auxiliary in the opposite sentence.

Function words, e.g. *of* or *which*, that are in one sentence but not the other were labelled CONTEXT or NONE depending on the human rater's judgment with respect to how strong the function word is implied by the other sentence. Function words play more of a syntactic role, i.e. they are more relevant in a context. If a function word is present in one sentence and not the other than it can only be linked to the opposite sentence via CONTEXT at best (or NONE).

All tokens (words/collocations and punctuation) must be explicitly matched (even if choosing the NONE matching).

Importantly, in greedy matching many-to-one relations are possible. In the example below, when matching *Duke* to a token in the other sentence it will be matched with *school*. Similarly, when *school* in the first sentence is matched it will be matched with *school* in the second sentence. Therefore, *Duke* and *school* in the first sentence will be matched to the same word, *school*, in the second sentence.

T1: Duke spokesman expressed concerns about the school's financial security.

T2: School representative expressed concerns about the university's financial security.

4. Optimal Annotation

The optimal matching strategy is inspired from optimal matching methods proposed for tasks where a set of items must be matched against another set while optimizing the overall matching score and not individual scores. The

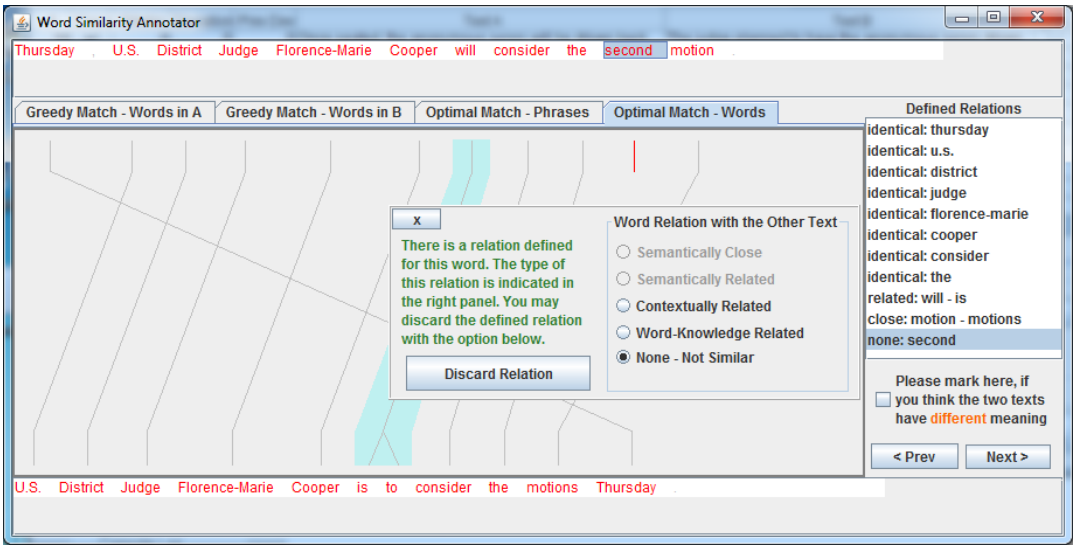

Figure 1. A snapshot of SIMILAT (SIMILar Annotation Tool).

overall matching score is the sum of individual scores for pairs of items, one from one set and the other item from the other set.

While in greedy matching the goal is for a target word to find a best matching word in the opposite sentence, in optimal matching the goal is to match items such that an overall optimal matching is achieved. Because it will be extremely time-consuming and error-prone to ask humans to fully emulate the optimal assignment algorithm, we simply asked them to pair words based on their full understanding of the two sentences. That is, given their reading of the two sentences judges were supposed to match the words that would make sense.

As opposed to greedy matching where one-to-many relations among words was possible, in optimal matching we strive for one-to-one matching.

An example of a pair of sentences where the greedy matching approach does not provide best overall, global match is given below.

T1: Duke spokesman expressed concerns about the school's financial security.

T2: School representative expressed concerns about the university's financial security.

In one matching, a target word, say *Duke* in the above example, can be greedily matched to the closest word in the other sentence, which is *university* (not *school*). In another matching, the target word *Duke* can be matched with the best matching word in the other sentence considering a more global assessment of both sentences. In our case, global matching would relate *Duke* with *school* and *school* in the first sentence with *university* in the second sentence.

Optimal matching involves matching words or phrases as best implied by the context of both sentences. Instead of focusing on a word, the focus is on finding the best match possible, which could be between two words, a word and a phrase, or two phrases. Optimal matching consists of two steps, as outlined below.

Step 1. Match chunks of the two sentences which are semantically equivalent beyond any doubts and whose equivalent meaning cannot be inferred from their words; that is, the meaning of these chunks could only grasped from the chunks as a whole; Examples of such semantically equivalent chunks/phrases are *give birth* and *have a child* or *have an offspring, living with the condition* and *suffering from a condition*.

Step 2. Eventually using information from Step 1, match individual words such that optimal matching is being achieved (at word-level). That is, a word should be matched against its best matching word as implied by the context of the two sentences and not necessarily its best individual match. For instance, a word should not be matched with an identical word in the opposite sentence if the context suggests the word should be matched to something else.

Examples of optimal matching are given below. The phrase *suffering from a condition* should be matched with the phrase *living with the condition* in the example below instead of just matching *suffering* with the word *condition* (based on individual similarities) or *suffering* with *living* (based on individual similarities and context).

T1: In Nigeria alone, the report said, as many as 1 million women may be living with the condition.

T2: In Nigeria alone, the report estimated that between 100,000 and 1 million girls and women are suffering from the condition.

For the pair of sentences below, the phrase *gives birth* and *has her first child* have the same meaning and therefore an optimal matching approach constrains the matching process to words within those phrases. That is, *birth* should only be matched to a word from the matching phrase *has her first child*.

T1: Crossing Jordan will be back in January after star Jill Hennessy gives birth.

T2: NBS also plans to shelve Crossing Jordan until January as star Jill Hennessy has her first child.

In this example below, no chunks should be selected as being equivalent because all chunks/phrases could be deemed similar (or not) based on their component words.

T1: The procedure is generally performed in the second or third trimester.
T2: The technique is used during the second and, occasionally, third trimester of pregnancy.

5. SIMILAT: The Semantic Annotation Tool

We have developed a tool to help our annotators easily annotate word-to-word relations. The annotation tool is called SIMILAT (SIMILarity Annotation Tool). A snapshot of the tool is shown in Figure 1.

The pair of two texts whose words are to be matched are shown at the top and bottom of SIMILAT's window. Below the text at the top, there are four tabs that support four different types of annotations: Greedy Match – Words in A, Greedy Match – Words in B, Optimal Match – Phrases, and Optimal Match – Words. Optimal Match – Phrases is a type of annotation that is currently under development and is not being described here. Greedy Match – Words in A allows the user to match one word at a time in the top text (called text A) to any word in the bottom text, called text B. This corresponds to the greedy annotation when target words are selected from text A. Similar, Greedy Match – Words in B allows the annotator to match one target word at a time in the bottom text to any word in text A. Optimal Match – Words facilitates optimal matching of words in which case any word in either text A or text B can be matched with a word and only one (or the whole context of the opposite sentence) or nothing in the other text. All the matchings can be done using the mouse by selecting the words to be matched and then choosing the type of relations from the pop-up menu: CLOSE, RELATED, CONTEXT, WORLD-KNOWLEDGE, and NONE. IDENTICAL matchings are automatically detected and shown in red.

As an annotator pairs certain words, they change their color to red to visually indicate they have been paired. The annotator must explicitly select a NONE relation for unmatched words so that they turn red. This assures that the annotator consider all the words explicitly. An annotator can move to the next pairs of sentences when all the words in the current pair are red, i.e. paired. An annotate pair is automatically saved when the annotator moves on to the next pair of sentences.

Besides providing the word-to-word similarity information, annotators were asked to judge whether the pair of sentences are indeed paraphrases or not. We wanted to compare such independent judgments with the original judgments provided by the MSRP designers. The annotation tool has a check button above the Prev and Next buttons at the bottom right corner of the SIMILAT's window that allows the annotators to specify whether they consider the two sentences to be in a paraphrase relation or not.

6. The SIMILAR Corpus

As we mentioned before, we selected a subset of the Microsoft Research Paraphrase (MSRP) corpus (Dolan, Quirk, and Brockett, 2004) to annotate. The MSR Paraphrase Corpus is the largest publicly available annotated paraphrase corpus which has been used in most of the recent studies that addressed the problem of paraphrase identification. The corpus consists of 5801 sentence pairs collected from newswire articles, 3900 of which were labelled as paraphrases by human annotators. The whole set is divided into a training subset (4076 sentences of which 2753 are true paraphrases) which we have used to determine the optimum threshold T, and a test subset (1725 pairs of which 1147 are true paraphrases) that is used to report the performance results.

There are several critiques about MSR corpus. First, MSR has too much word overlap (spawning from the way they collected the data set) and less syntactic diversity. Therefore, the corpus cannot be used to learn paraphrase syntactic patterns (Zhang and Patrick 2005; Weeds 2005). It should be noted that the lexical overlap is recognized by the creators of the corpus (Dolan and Brockett 2005) which indicate a .70 measure of overlap (of an unspecified form). The T-F split in both training and testing is quite similar though (67-33%).

Second, the annotations by humans were made on slightly modified sentences which are different from the original sentences publicly released. For instance, humans were asked to ignore all numbers and simply replace them with a generic token, e.g. MONEY for monetary values, and make judgments accordingly. This discrepancy between what humans used and what systems take as input complicates the task as some decisions are counterintuitive. For instance, the pair below was judged as a paraphrase although the percentages as well as the indices (*Standard & Poor* versus *Nasdaq*) are quite different.

T1: The broader Standard & Poor's 500 Index .SPX gained 3 points, or 0.39 percent, at 924.

T2: The technology-laced Nasdaq Composite Index < :IXIC > rose 6 points, or 0.41 percent, to 1,498.

Nevertheless, the MSRP corpus is the largest available and most widely used.

We annotated 700 pairs of sentences from the MSRP corpus which consists of 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 most of the semantic similarity metrics we used to derive semantic similarity scores with in order to relate to the human annotations only 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. Our own judgments yielded 63% (442) TRUE paraphrases for an overall agreement rate between our annotations and the MSRP annotations (both TRUE and FALSE paraphrases) of 75.7%. We simply instructed our judges to use their own judgment with respect to whether the two sentences mean

	Close	Related	Context	World Knowledge
Resnick	0.718	0.465	0.348	0.340
Leacock-Chodorow	0.862	0.639	0.596	0.499
Jiang and Conrath	0.774	0.268	0.190	0.191
Path	0.757	0.358	0.298	0.222
Lin	0.893	0.588	0.506	0.446
Wu and Palmer	0.886	0.701	0.605	0.578
LSA	0.292	0.228	0.136	0.204

Table 3. Average scores for each type of relation and each word-to-word similarity metric (all greedily matched pairs of words were included; from Text 1 to Text 2 and from Text 2 to Text 1).

	Close	Related	Context	World Knowledge
Resnick	0.702	0.5	0.33	0.249
Leacock-Chodorow	0.844	0.678	0.571	0.439
Jiang and Conrath	0.735	0.314	0.17	0.163
Path	0.728	0.412	0.268	0.188
Lin	0.869	0.632	0.449	0.339
Wu and Palmer	0.871	0.733	0.601	0.495
LSA	0.278	0.217	0.127	0.132

Table 4. Average scores for each type of relation and each word-to-word similarity metric for the optimally matched pairs for words.

	Close	Related	Context	World Knowledge
Resnick	0.375 (334/890)	0.634 (788/1242)	0.544 (241/443)	0.617 (169/274)
Leacock-Chodorow	0.336 (299/890)	0.559 (694/1242)	0.372 (165/443)	0.529 (145/274)
Jiang and Conrath	0.384 (342/890)	0.597 (742/1242)	0.424 (188/443)	0.693 (190/274)
Path	0.336 (299/890)	0.559 (694/1242)	0.372 (165/443)	0.529 (145/274)
Lin	0.416 (370/890)	0.648 (805/1242)	0.535 (237/443)	0.748 (205/274)
Wu and Palmer	0.336 (299/890)	0.561 (697/1242)	0.379 (168/443)	0.529 (145/274)
LSA	0.334 (297/890)	0.553 (687/1242)	0.381 (169/443)	0.507 (139/274)

Table 5. Percentage and raw numbers in parenthesis of pairs of greedily matched words for which the word-to-word semantic similarity metrics could not provide a score indicating their limitation.

	Close	Related	Context	World Knowledge
Resnick	0.383 (151/394)	0.619 (234/378)	0.58 (138/238)	0.721 (49/68)
Leacock-Chodorow	0.33 (130/394)	0.548 (207/378)	0.45 (107/238)	0.647 (44/68)
Jiang and Conrath	0.376 (148/394)	0.579 (219/378)	0.542 (129/238)	0.809 (55/68)
Path	0.33 (130/394)	0.548 (207/378)	0.450 (107/238)	0.647 (44/68)
Lin	0.414 (163/394)	0.614 (232/378)	0.630 (150/238)	0.853 (58/68)
Wu and Palmer	0.33 (130/394)	0.55 (208/378)	0.454 (108/238)	0.647 (44/68)
LSA	0.322 (127/394)	0.532 (201/378)	0.471 (112/238)	0.515 (35/68)

Table 6. Percentage and raw numbers in parenthesis of pairs of optimally matched words for which the word-to-word semantic similarity metrics could not provide a score indicating their limitation.

the same thing or not. 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.

We have annotated so far 700 pairs. The 700 pairs were annotated by 6 different judges each annotating an equal, separate subset. As of this writing, a second judge annotates the same subset and we will be able to report inter-judge agreement. On a trial exercise of 100 pairs, inter-judge reliability was 63% at individual relation level.

Our effort resulted in a total of 12,560 relations of which 8,346 were IDENTICAL matches, 2849 relations detected greedily (890 CLOSE relations, 1242 RELATED relations, 443 CONTEXT relations, 274 KNOWLEDGE relations) and 1966 words were unmatched (a NONE type of relation was assigned to these words). For the optimum annotation, 15,692 relations were identified of which 8,046 were IDENTICAL and the judges identified 1,078 relations (394 CLOSE relations, 378 RELATED relations, 238 CONTEXT relations, 68 KNOWLEDGE relations) and 4,306 words were non matched.

We compared the human annotations with results

obtained with the word-to-word semantic similarity measures in the WordNet Similarity library (Pedersen, Patwardhan, and Michelizzi, 2004) as well as using LSA (Landauer et al., 2007).

We used the following similarity measures implemented in the WordNet::Similarity package and described in Pedersen, Patwardhan, and Michelizzi (2004): LCH (Leacock & Chodorow, 1998), RESNIK (Resnik, 1995), JIANG and CONRATH (Jiang & Conrath, 1997), LIN (Lin, 1998), PATH (Pedersen, Patwardhan, and Michelizzi, 2004) and WUP (Wu & Palmer, 1994). The WordNet-based similarity metrics require words with senses (i.e. concepts in WordNet; Miller, 1995) as input. We have experimented with all combinations of senses. We also used LSA as a word-to-word similarity metric. The LSA vectors were derived from a large collection of texts (the TASA corpus; Zeno et al., 1995).

The results are summarized in Tables 3-6. To obtain the results we took all matched words by humans and computed word-to-word similarity scores with each of the word-to-word semantic similarity metrics (shown in the first column) .Table 3 presents the average scores for all the similar words matched by the human annotators per the type of qualitative similarity relation identified by the annotators. Table 3 presents results for similar words that were greedily matched while Table 4 for words optimally matched. Table 3 combined the results for the greedy annotations in both directions: matching target words from text A to words in text B and also matching target words from text B to words in text A. From both tables 3 and 4 we can clearly see that the averages for each type of relations are very different with few exceptions. For instance the Jiang and Conrath and the LSA cannot distinguish between CONTEXT and KNOWLEDGE types of relations when optimally matched. LSA yields very close averages for RELATED and KNOWLEDGE types of relations when greedily matched. Resnick also has problems separating the CONTEXT from KNOWLEDGE word matchings when greedily matched as the corresponding averages are very close.

When analyzing the results in Tables 5 and 6, which represent the percentages of pairs of words by annotators for which the word-to-word semantic similarity metrics could not provide a score (i.e. misses), we realized that LSA is the most robust as it has least misses. The other measures are constraint to only content words or only certain types of words, e.g. nouns or verbs. LSA could compute the similarity between a pronoun and noun, for instance, while any of the WordNet Similarity metrics cannot. The Lin measure yields the most misses.

7. Further Work

We plan to continue our work presented in this paper along several lines of future research. First, we would like to annotate more data to have a larger annotated corpus. Furthermore, we would like to add another level of annotation in which we indicate phrases that are semantically equivalent without the need to matched particular words within those phrases. Such examples of equivalent phrases which do not need to be decomposed further into word-level matchings are “giving birth” and “have an offspring”. Second, we plan to use the greedily matched pairs and the optimally matched pairs by human annotators in automated methods and compare the results thus obtained with the fully greedy and automated methods. Finally, we would like to propose a qualitative model of word-level semantic similarity.

8. Conclusion

We have described in this paper a novel protocol to annotate texts with qualitative judgments of word-level similarity. A greedy and optimal annotation strategy was developed and implemented. The word-to-word annotations by human judges were related to quantitative scores of similarity generated by a set of WordNet-based similarity metrics and LSA. The comparison revealed the strengths and weaknesses of these metrics which in turn has important implications for future developments of text-to-text similarity methods and other methods that will include the word-to-word similarity metrics.

9. 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.

10. References

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* **3** (4–5): *pp.* 993–1022.

Corley, C. and Mihalcea, R. 2005. Measures of Text Semantic Similarity, in *Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence*, Ann Arbor, MI, June 2005

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

Dasgupta, D., Niño, F., Garrett, D., Chaudhuri, K., Medapati, S., Kaushal, A., Simien, J. 2009. A multi-objective evolutionary algorithm for the task based sailor assignment problem. GECCO 2009: 1475-1482.

Dawes, M. 2011. *The Optimal Assignment Problem*, Course notes, University of Western Ontario. (accessed online in December 2011)

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.

Graesser, A.; Hu, X.; and McNamara, D. 2005. Computerized learning environments that incoporate research in discourse psychology, cognitive science, and computational linguistics. In Healy, A., ed., *Experimental Cognitive Psychology and its*

Applications, 59–72. Washington, D.C. American Psychological Association.

Graesser, A.; Olney, A.; Hayes, B. C.; and Chipman, P. 2005. Autotutor: A cognitive system that simulates a tutor that facilitates learning through mixed-initiative dialogue. In *Cognitive Systems: Human Cognitive Models in System Design*. Mahwah: Erlbaum.

Ibrahim, A., Katz, B., and Lin, J. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceeding of the Second International Workshop on Paraphrasing*, (ACL 2003).

Iordanskaja, L., Kittredge, R., and Polgere, A. 1991. Natural Language Generation in Artificial Intelligence and Computational Linguistics. Lexical selection and paraphrase in a meaning-text generation model, Kluwer Academic.

Jiang, J.J. & Conrath, D.W. 1997. Semantic Similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*.

Kuhn, H.W. 1955. "The Hungarian Method for the assignment problem", *Naval Research Logistics Quarterly*, 2:83–97, 1955. Kuhn's original publication.

Landauer, T.K..; McNamara, D.S.; Dennis, S.; and Kintsch, W. 2007. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Leacock, C.; and Chodorow, M. 1998. Chapter: Combining local context and WordNet sense similarity for word sense identification. *WordNet, An Electronic Lexical Database.* The MIT Press.

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.

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

Miller, G. 1995. WordNet: A Lexical Database of English. *Communications of the ACM*, v.38 n.11, p.39-41.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. WordNet::Similarity – Measuring the Relatedness of Concepts. In *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2004.*

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

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., McCarthy, P. M., Lintean, M., McNamara, D. S., and Graesser, A. C. 2008. Paraphrase Identification with Lexico-Syntactic Graph Subsumption. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS-2008)*.

Rus, V., Lintean, M., Azevedo, R. (2009). Automatic Detection of Student Mental Models During Prior Knowledge Activation in MetaTutor. Proceedings of the 2nd International Conference on Educational Data Mining. Cordoba, Spain.

Rus, V., Nan, X., Shiva, S., & Chen, Y. 2009. Clustering of Defect Reports Using Graph Partitioning Algorithms, *Proceedings of the 20th International Conference on Software and Knowledge Engineering*, July 2-4, 2009, Boston, MA. Rus, V., McCarthy, P. M., Lintean, M., McNamara, D. S., and Graesser, A. C. 2008. Paraphrase Identification with Lexico-Syntactic Graph Subsumption. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS-2008)*.

Rus, V. and Lintean (2012). A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics. Proceedings of the International Conference on Intelligent Tutoring Systems. Crete, Greece.

Salton, A. G., Wong, and C. S. Yang. 1975. "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, nr. 11, pages 613–620.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A. M., & Rose, C. 2007. When are tutorial dialogues more effective than reading? *Cognitive Science, 31*, 3-62.

Weeds, J., Weir, D., & Keller, B. 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.

Wu, Z.; and Palmer, M.S. 1994. Verb Semantics and Lexical Selection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. 1995. *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

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