SEMILAR: A Semantic Similarity Toolkit For Assessing Students' Natural Language Inputs

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

Rajendra Banjade
Department of Computer Science
The University of Memphis
Memphis, TN 28152
rbanjade@memphis.edu

Mihai Lintean
Department of Computer Science
The University of Memphis
Memphis, TN 38152
mclinten@memphis.edu

Nobal Niraula
Department of Computer Science
The University of Memphis
Memphis, TN 38152
nbnraula@memphis.edu

Dan Stefanescu
Department of Computer Science
The University of Memphis
Memphis, TN 38152
dnstfnscu@memphis.edu

ABSTRACT

We present in this demo SEMILAR, a SEMantic similarity toolkit. SEMILAR includes offers in one software environment several broad categories of semantic similarity methods: vectorial methods including Latent Semantic Analysis, probabilistic methods such as Latent Dirichlet Allocation, greedy lexical matching methods, optimal lexico-syntactic matching methods based on word-to-word similarities and syntactic dependencies with negation handling, kernel based methods, and some others. We will demonstrate during this demo presentation the efficacy of using SEMILAR to investigate and tune assessment algorithms for evaluating students' natural language input based on data from the DeepTutor computer tutor.

Keywords

Natural language student inputs, assessment, conversational tutors.

1. INTRODUCTION

In dialogue-based Intelligent Tutoring Systems (ITS; Rus, D'Mello, Hu, & Graesser, in press; Evens & Michael, 2005), it is important to understand students' natural language responses. Accurate assessment of students' responses enables the building of accurate student models for both cognition and affect. An accurate student model in turn affects the quality of tutor's feedback (Rus & Lintean, 2012). In general, accurate student models lead to improved macro- and micro-adaptivity in ITSs which is needed for effective tutoring (Rus, D'Mello, Hu, & Graesser, in press).

There are at least two different types of natural language assessments in conversational ITSs. First, there is need for advanced natural language algorithms to interpret the meaning of students' natural language contributions at each turn in the dialogue. The student responses in the middle of the dialogue tend to be short, i.e. the length of a sentence or less. There is also a need to assess the more comprehensive, essay-type answers that students are required to provide immediately after being prompted to solve a problem. These essay-type answers can be a paragraph long or even longer depending on the task and target domain.

One approach to assessing students' responses is to compute how similar the responses are to benchmark solutions provided by experts (Rus & Graesser, 2006). That is, *semantic similarity* is the underlying principle for computing the meaning of student contributions in many of today's state-of-the-art conversational ITSs and in other mainstream natural language processing applications such as Question Answering or Paraphrase Identification. The alternative approach to natural language understanding, called true understanding, is impractical as it requires world knowledge which is an intractable problem in Artificial Intelligence.

As already mentioned, in the semantic similarity approach a student contribution is assessed in terms of its similarity to an expert answer. The expert answer is deemed correct. Therefore, a student contribution is deemed correct if it is semantically similar to the expert answer (and incorrect otherwise).

Below, we show an example of a real student response from an ITS and the corresponding expert-answer as authored by an expert.

Student Response: *An object that has a zero force acting on it will have zero acceleration.*

Expert Answer: *If an object moves with a constant velocity, the net force on the object is zero.*

The student response above is deemed correct as it is semantically similar to the expert answer. In general, the student response is deemed incorrect if it is not semantically similar enough to the expert response. More nuanced assessments can be made (e.g., partially correct or partially correct and partially incorrect at the same time).

Researchers have been proposing various methods to assess the semantic similarity of texts, in particular sentences (Corley and Mihalcea, 2005; Fernando & Stevenson, 2008; Rus, Lintean, Graesser, and McNamara 2009). However, there is no software library or toolkit that would allow for a fair comparison and investigation of the various methods. Furthermore, there is no one-stop-shop kind of environment to explore semantic similarity methods at various levels of granularity: word-to-word, sentence-to-sentence, paragraph-to-paragraph, or document-to-document

similarity. Furthermore, mixed combinations of similarity could be imagined such as examining how similar a summary paragraph is to a document (useful in summarization).

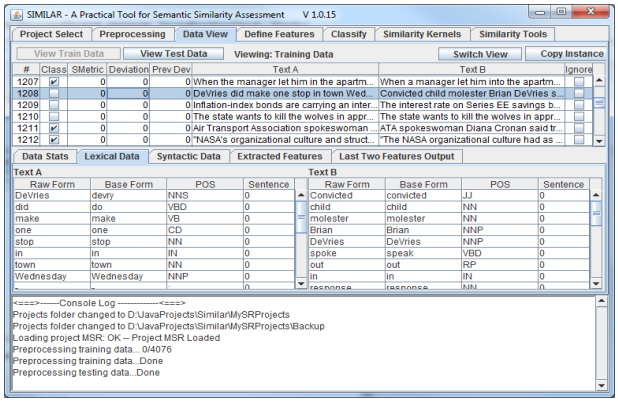

Figure 1. Snaphsot of SEMILAR (Data View Pane).

Given the importance of assessing students' natural language inputs for building accurate student models, there is an acute need for such a software environment that would allow for a systematic and fair comparison of the various semantic similarity methods to assess students' natural language inputs.

The proposed SEMILAR (SEMantic simILARity) toolkit address this need by offering a java library as well as a GUI-based Java application that integrates a myriad of semantic similarity methods for tuning and optimizing the parameters of such methods for the student assessment task in conversational ITSs.

2. SEMILAR: THE SEMANTIC SIMILARITY TOOLKIT

The authors of the SEMILAR toolkit have been involved in assessing the semantic similarity of texts for more than a decade.

During this time, they have conducted a careful requirements analysis for an integrated software toolkit to be used for semantic similarity assessment. The result of this effort is the prototype presented here.

The SEMILAR toolkit includes the following components: project management; data view/browsing/visualization; textual preprocessing (e.g., tokenization, lemmatization/stemming, collocation identification, part-of-speech tagging, phrase or dependency parsing, etc.), semantic similarity methods, classification components (naïve Bayes, Decision Trees, Support Vector Machines, and Neural Network), kernel-based methods (sequence kernels, word sequence kernels, and tree kernels; as of this writing, we are still implementing several other tree kernel methods); debugging and testing facilities for model selection; and annotation components (allows domain expert to manually annotate texts with semantic relations using GUI-based facilities). For space reasons, we will only detail next the core component that includes the text-to-text similarity algorithms available in SEMILAR.

We briefly present core methods available as of this writing:

- a greedy method based on word-to-word similarity measures
- an optimal matching solution 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;
- 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), typically used in textual entailment which as a text-to-text semantic relation;
- a method in which similarities among all pairs of words are taken into account for computing the similarity of two texts. A similarity matrix operator W that contains word-to-word similarities between any two words is used;
- a weighted-LSA (wLSA) method for semantic similarity based on Latent Semantic Analysis. 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. A combination of 3 local weights and 3 global weights are available.
- A set of similarity measures based on the unsupervised method Latent Dirichlet Allocation. 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).
- The Quadratic Assignment Problem (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 formulation of the QAP problem best fits the purpose of semantic similarity. The QAP method provides best accuracy results (=77.6%) that rival the best reported results so far (Madnani, Tetreault & Chodorow, 2012).

3. ACKNOWLEDGMENTS

This research was supported in part by Institute for Education Sciences under awards R305A100875.

4. REFERENCES

[1] Evens, M., and Michael, J. 2005. One-on-one Tutoring by Humans and Machines. Mahwah, NJ: Lawrence Erlbaum Associates.

[2] Graesser, A. C., Rus., V., D'Mello, S., K., & Jackson, G. T. (2008). AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. Schraw (Eds.), Current perspectives on cognition, learning and instruction: Recent innovations in educational technology that facilitate student learning (pp. 95–125). Information Age Publishing.

[3] 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 Innovative Use of Natural Language Processing for Building Educational Applica-tions, NAACL-HLT 2012, Montreal, Canada, June 7-8, 2012.