Chapter 7
Text-to-Text Similarity of Sentences

Vasile Rus
The University of Memphis, USA

Mihai Lintean
The University of Memphis, USA

Arthur C. Graesser
The University of Memphis, USA

Danielle S. McNamara
Arizona State University, USA

ABSTRACT
Assessing the semantic similarity between two texts is a central task in many applications, including summarization, intelligent tutoring systems, and software testing. Similarity of texts is typically explored at the level of word, sentence, paragraph, and document. The similarity can be defined quantitatively (e.g. in the form of a normalized value between 0 and 1) and qualitatively in the form of semantic relations such as elaboration, entailment, or paraphrase. In this chapter, we focus first on measuring quantitatively and then on detecting qualitatively sentence-level text-to-text semantic relations. A generic approach that relies on word-to-word similarity measures is presented as well as experiments and results obtained with various instantiations of the approach. In addition, we provide results of a study on the role of weighting in Latent Semantic Analysis, a statistical technique to assess similarity of texts. The results were obtained on two data sets: a standard data set on sentence-level paraphrase detection and a data set from an intelligent tutoring system.

INTRODUCTION
Computational approaches to language understanding can be classified into two major categories: true-understanding and text-to-text similarity. In true understanding, the goal is to map language statements onto a deep semantic representation that relate language constructs to world and domain knowledge. Current state-of-the-art approaches that fall into this true-understanding category offer adequate solutions only in very limited contexts (i.e. toy-domains) lacking scalability and thus having limited use in real world applications such as summarization or intelligent tutoring systems.
Text-to-text similarity approaches (T2T) to text semantic analysis avoid the hard task of true understanding by defining the meaning of a text based on its similarity to other texts, whose meaning is assumed to be known. Such methods are called benchmarking methods as they rely on a benchmark text, analyzed by experts, to indentify the meaning of new, unseen texts. We adopt in this chapter a T2T approach to semantic text analysis.

In particular, we focus on the task of quantifying how similar two texts are, and based on this, we then decide whether they are similar enough to be considered a paraphrase or not. An example of two texts, a textbase (T) and student paraphrase (SP; reproduced as typed by the student in iSTART, an intelligent tutoring system; McNamara, Levinstein, & Boonthum, 2004), is provided below (from the User Language Paraphrase Challenge; McCarthy & McNamara, 2008):

- **T**: During vigorous exercise, the heat generated by working muscles can increase total heat production in the body markedly.
- **SP**: alot of excercise can make your body warmer.

Human judges deemed the T and SP in this example to be similar (i.e. in a paraphrase relationship).

We present in this chapter two categories of approaches to the task of sentence-level paraphrase identification: knowledge-based and statistical-based. A generic approach that relies on knowledge-based word-to-word similarity measures is discussed. In addition, we present a generic approach based on Latent Semantic Analysis (LSA; Landauer et al., 2007), a statistical technique to assess similarity of texts, which is used in combination with several weighting schemes to address the task of paraphrase identification. These approaches were tested on two data sets: the Microsoft Research Paraphrase corpus (MSRP; Dolan, Quirk, & Brockett, 2004), a standard data set on sentence-level paraphrase detection, and a data set from the intelligent tutoring system iSTART (McCarthy & McNamara, 2008).

**BACKGROUND**

In this section, we present background information related to word-level similarity measures, as they form the foundation of methods we propose.

There are two main groups of word-level similarity techniques: knowledge-based and statistical. In the knowledge-based category, the lexical database WordNet is used as a knowledge base (Miller, 1995). WordNet groups words with same meaning into synsets (i.e. synonymous sets). Each synset defines a concept (i.e. a uniquely identified meaning). A word can belong to more than one synset in cases where the word is polysemous (i.e. it has many senses). WordNet contains only content words: nouns, verbs, adjectives, and adverbs. It should be noted that WordNet simply offers a glosary of possible senses for words. Identifying the exact meaning (out of many) of a word according to WordNet is equivalent to identifying the synset that best captures the meaning of the word given its context in a particular text fragment. That is, the meaning of the word is entailed by the company it keeps in a text fragment. The task of identifying the correct sense given the context is called word sense disambiguation, one of the most difficult tasks in natural language processing. Text-to-text similarity methods that rely on WordNet-based word-to-word similarity measures need word sense disambiguation in order to be used.

There are nearly a dozen WordNet-based similarity measures available (Pedersen, Patwardhan, & Michelizzi, 2004). These measures are further divided into two groups: similarity measures and relatedness measures. The similarity measures are limited to within-category concepts (noun-noun or verb-verb) and usually they work only for nouns and verbs. The text relatedness measures on the