KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2007
Cluj-Napoca (Romania), June 6–8, 2007, pp. 9–10

SEMANTIC SIMILARITY KNOWLEDGE AND ITS APPLICATIONS

DIANA INKPEN

Abstract

Semantic relatedness refers to the degree to which two concepts or words are related. Humans are able to easily judge if a pair of words are related in some way. For example, most would agree that "apple" and "orange" are more related than are "apple" and "toothbrush". Semantic similarity is a subset of semantic relatedness.

In this talk I will present several methods for computing the similarity of two words, following two directions: dictionary-based methods that use WordNet, Roget's thesaurus, or other resources; and corpus-based methods that use frequencies of co-occurrence in corpora (cosine method, latent semantic indexing, mutual information, etc). I will present several applications of word similarity knowledge: detecting words that do not fit into their context (real-word error correction), detecting speech recognition errors, solving TOEFL-style synonym questions, and synonym choice in context (for writing aid tools).

I will also present a method for computing the similarity of two texts, based on the similarities of their words. Applications of text similarity knowledge include: designing exercises for second language-learning, acquisition of domain-specific corpora, information retrieval, and text categorization.

At the end, I will present cross-language extensions of the methods for similarity of words and texts.

Biography

Dr. Diana Inkpen is an Assistant Professor of Computer Science at the School of Information Technology and Engineering, University of Ottawa since July 2003. She obtained her doctorate from the University of Toronto, Department of Computer Science. She has a Masters in Computer Science and Engineering from the Technical University of Cluj-Napoca, Romania. Her research projects and

2000 *Mathematics Subject Classification.* 68T35, 68T50, 91F20.

publications are in the areas of Computational Linguistics and Artificial Intelligence, more specifically: Information Retrieval, Information Extraction, Natural Language Understanding, Natural Language Generation, Speech Processing, and Intelligent Agents for the Semantic Web.