論文の内容の要旨

論文題目：A Study on Attributional and Relational Similarity between Word Pairs on the Web

(ウェブ上での単語対間の属性類似性と関係類似性に関する研究)

氏名：ボッレーガラ ダヌシカ タルパティ

Similarity is a fundamental concept that extends across numerous fields such as artificial intelligence, natural language processing, cognitive science and psychology. Similarity provides the basis for learning, generalization and recognition. Similarity can be broadly divided into two types: semantic (or attributional) similarity, and relational similarity. Attributional similarity is the correspondence between the attributes of two objects. If two objects have identical or close attributes, then those two objects are considered attributionally similar. For example, the two concepts, *Jaguar* and *cat*, both have an identical set of attributes: both are mammals, have four legs, and carnivorous animals. Consequently, the two words Jaguar and cat show a high degree of attributional similarity. On the other hand, relational similarity is the correspondence between the implicit semantic relations that exist between two pairs of words. For example, consider the two word-pairs (*ostrich*, *bird*) and (*lion*, *cat*). Ostrich is a large bird and lion is a large cat. The implicitly stated semantic relation *is a large* holds between the two words in each word-pair. Therefore, those two word-pairs are considered relationally similar. Typically, word analogies show a high degree of relational similarity.

This thesis addresses the problem of measuring both attributional and relational similarity between words or pairs of words from the web. In Chapter 1, I define the two types of similarity in detail and present the overall structure of the thesis. Chapter 2 presents a supervised approach to measure the semantic similarity between two words using a web search engine. The proposed method reports a high correlation with human ratings in a benchmark dataset for semantic similarity. The proposed semantic similarity is used in a community clustering task and a word sense disambiguation task. Chapter 3 studies the problem of relational similarity. To represent the implicitly stated semantic relations between two words, I extract lexical patterns from the snippets retrieved from a web search engine for the two words. Then the extracted patterns are clustered using distributional similarity to identify the different patterns that describe a particular semantic relation. Finally, machine learning approaches are used to compute the relational similarity between two given word-pairs using the lexical patterns extracted for each word-pair. I experiment with support vector machines and information theoretic metric learning approach to learn a relational similarity measure.

The second half of this thesis describes the applications of semantic and relational similarity. As a working problem, I concentrate on personal name disambiguation on the web. A name of a person can be ambiguous on the web because of two main reasons. First, different people can share the same name (namesake disambiguation problem). Second, a single individual can have multiple aliases on the web (alias detection problem). Chapter 4 analyzes the namesake disambiguation problem, whereas, Chapter 5 focuses on the alias detection problem. I propose fully automatic methods to solve both these problems with high accuracy. In Chapter 6, I present a preliminary work on personal attribute extraction from the web.

In Chapter 7, I present a relational model of semantic similarity that connects relational and attributional similarity measures that were introduced in the thesis. In contrast to the feature model of semantic similarity, which models objects using their attributes, the relational model attempts to compute the semantic similarity between two given words directly using the numerous semantic relations that hold between the two words. I conclude this thesis with a description of potential future work in web-based similarity measurement.