

Research Statement

Xiaohua (Davis) Zhou

College of Information Science and Technology, Drexel University, Philadelphia, PA 19104

Overview

My research theme has been the development of theories and tools for effective ways of utilizing semantic knowledge for information retrieval (IR) and text mining (TM). As shown in Figure 1, centering on semantic-based language modeling (SBLM), my dissertation and other research projects work around the following five questions: (1) how to *represent* semantic knowledge, (2) how to *acquire* semantic knowledge, (3) how to *incorporate* semantic knowledge into statistical language models, (4) how to *apply* SBLM to text retrieval, question answering, document clustering, and document classification, and (5) how to *evaluate* the effectiveness of SBLM and its applications in different domains including biomedical literature, healthcare, news, and web. Based on these research questions, I have authored or co-authored four papers in journals including the *IEEE Transaction on Knowledge and Data Engineering* and thirteen papers in some of the most competitive conferences such as *SIGIR*, *CIKM*, *IJCAI* and *ICDM* over the past four years. I also implemented a comprehensive IR and TM research tool, the *Dragon Toolkit*, as a result of my research. Currently, the toolkit is being shared as a workbench by more than 500 researchers in the areas of IR and TM since its first release in April, 2007. The following are some brief descriptions of my research projects.

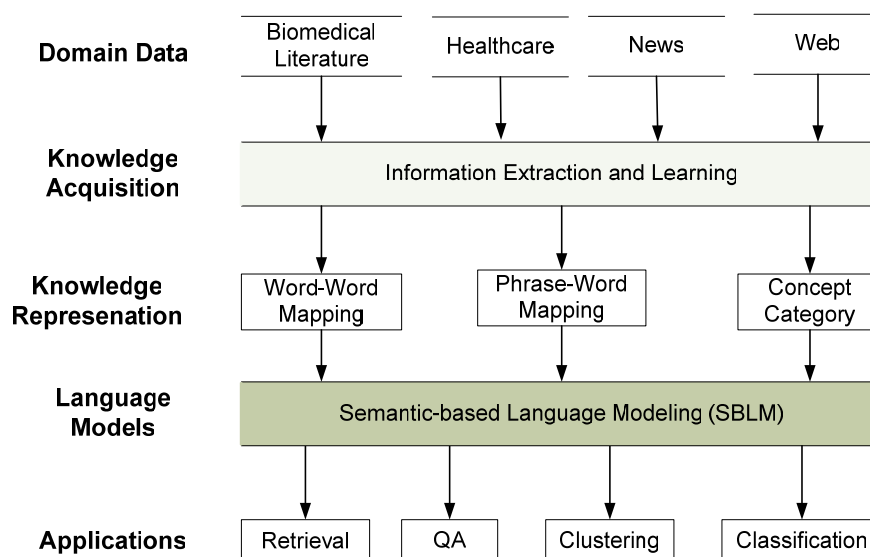


Figure 1. Research Framework. The implementation of this framework, the *Dragon Toolkit* is publicly available at <http://www.dragontoolkit.org>.

Research Projects

Text Retrieval Recognizing that biomedical literature contains a large number of terms such as genes, proteins, cells and diseases which often have many synonyms as well as various inter-relationships, I developed semantic-based language models to solve the problem. A semantic vector for each term

represents the probabilities the term could map to other terms in the vocabulary (e.g. UMLS Metathesaurus). The language models smoothed by the semantic vector significantly improved the retrieval results. Furthermore, the term-term mapping outperformed the word-word mapping due to the ambiguity of individual words. The mapping algorithm and the new language model were published in *SIGIR 2006*. Later, I extended the semantic-based language models to public domains such as news collections. Since there is no domain ontology available in public domains, multiword phrases are extracted and phrase-word mapping are used instead. A comprehensive version of SBLM for text retrieval was published in the *IEEE Transaction on Knowledge and Data Engineering (TKDE) 2007*.

Text Categorization There are two challenging problems in text classification and clustering. One is the highly skewed data, i.e. some classes may contain few documents. The other is the availability of labeled training data for text classification. Thus, building an accurate classifier with small training datasets becomes very useful. It is often difficult for traditional language models to cope with classes with few documents. Instead, I used semantic mapping vectors to smooth class (cluster) language models. The new approach improved both text clustering and classification in two aforementioned situations for biomedical literature and news collections. Two papers in *IJCAT 2007* and *ICDM 2006* described semantic-based language modeling approaches to agglomerative clustering and partitional clustering, respectively. A paper under the review of *ACM Transactions on Information Systems* presents the new classification method and the comprehensive comparisons among concept-word mapping, phrase-word mapping and word-word mapping.

Question Answering In two recent ongoing projects, I am utilizing web search engines for entity labeling as well as factoid question answering. In the first project, an entity is submitted to a search engine and the returned top snippets are used to build a context vector, which is further labeled by a trained SVM classifier. This approach does not require external language resources, but achieves quite promising results in open domains (demo: <http://www.dragontoolkit.org/cptc.asp>). The result is utilized as a candidate answer filtering component by the second QA project. In the QA project (demo: <http://www.dragontoolkit.org/qa.asp>), I apply SBLM to question classification as well as candidate answer re-ranking. Since both a question and the local context of a candidate answer are quite short, the SBLM performs very effectively. A comparison to the state-of-the-art ARANEA system shows great potential of SBLM approaches to the factoid question answering.

Future Research

My future research will continue to broaden and deepen studies in information retrieval and text mining. I plan to pursue the following stream of research in my future faculty career. First, I will extend my research framework to the domain of digital library. My favorite research topics include reference question classification and answering (e.g. IPL project), domain-specific search, and semantic annotation. Second, I have particular interest in contextual retrieval. Current search engines view queries independent of each other. In fact, like berry-picking, users often clarify their information needs after reformulating the original query for a couple of times. The detection of such dynamic search context is very useful, but quite challenging with so short queries. Semantic-based language modeling may help solve this problem. Third, I will continue my effort on web-based question answering. The focus will shift from simple factoid questions to more complex questions.