

Research Statement

Jing JIANG

School of Information Systems, Singapore Management University

Tel: (65) 6828-0785; Email: jingjiang@smu.edu.sg

Updated on 15 February, 2009

Background

With the explosion of the amount of textual data in the information age, there is an urgent need for powerful text information management tools to help people quickly find, extract, digest and summarize information stored in all kinds of text. For example, Web search engines are effective tools to help us find relevant Web pages, which mainly consist of text. Employees of an organization need special enterprise search tools to handle intranet pages, email archives, shared text files, etc. Recent growth of user-generated content on the Web such as product reviews and blogs also requires new tools to extract information such as consumer sentiments from these unconventional data sources. A common challenge in improving the effectiveness of existing tools as well as building novel tools is to develop adaptable and scalable language technologies that enable computers to understand text. The goal of my research is to develop general and effective techniques for understanding natural language and extracting useful information from text, and to apply these techniques to critical, real-world information management problems.

Current Research

Domain Adaptation Machine learning supplemented by linguistic and domain knowledge has proven to be the most effective approach to understanding human language and making use of textual data. However, current machine learning approach is mostly based on supervised learning, which works well only when sufficient human-annotated text is available for training. This rarely happens in large-scale applications due to the labor-intensive nature of human annotation. A major line of my research is to address this bottleneck problem with insufficient training text. In particular, I am interested in the setting where we have no or a little annotated text for the problem at hand but a large amount of annotated text in some related domains. For example, to extract information from online blog entries, we may “borrow” news articles, which have been extensively annotated and studied in the past. Despite the importance of this *domain adaptation* problem, we still have limited understanding of it so far, and it is very challenging to formally formulate the problem and to develop general and effective solutions. I have developed two frameworks to address domain adaptation, one based on instance weighting and the other on feature selection [1, 2]. In both frameworks, adaptation is achieved through modifying the standard risk minimization objective function. The two frameworks have been evaluated on a number of text mining tasks such as named entity recognition and spam

filtering, and outperformed regular supervised and semi-supervised learning methods. Together with my collaborators, I have also explored ensemble models to tackle the domain adaptation problem [3].

Information Extraction Information extraction aims at finding specific chunks of information such as names of people and organizations and relations between these named entities from text. It is an important text mining problem that has many applications in question answering, data mining, etc. I have looked at two aspects of the limitations of current information extraction techniques. First, I studied the domain adaptation problem for named entity recognition. Using feature selection based domain adaptation techniques, we were able to significantly improve the performance [4, 1]. Second, I studied the feature engineering problem for extracting semantic relations between entities from text [5]. With the feature space we defined, feature selection for relation extraction can be done in a more systematic manner.

Transfer Learning for Relation Extraction Transfer learning is a more general problem than domain adaptation, in which annotated text can be borrow from not only a related domain but also a related task. I am currently looking into the problem of applying transfer learning techniques for relation extraction, and in particular, how annotated text for one type of relations between two entities (e.g. employment relation) can help the extraction of another type of relations (e.g. membership relation). Our preliminary results suggest that such knowledge transfer is indeed possible, especially with the guidance of expert knowledge.

Future Research

I plan to continue my research in the aforementioned directions with the ultimate goal of making text mining tools readily adaptable to new domains and new tasks. A major challenge in adapting text mining tools to new domains is to gather domain knowledge. I want to explore the direction of exploiting the human knowledge already accumulated on the Web in user-generated content such as Wikipedia, social bookmarks and blogs. The challenge then becomes how to construct domain ontology from these existing but hidden knowledge bases.

Another general direction I plan to explore is to develop general methods to automatically generate integrated summaries of search results. Current search results are presented as ranked lists of documents, which are not appealing to the user. One solution is to automatically generate structured text summaries for different types of user queries, using adaptive information extraction techniques. If information extraction can be made easily adaptable, for different types of user queries, we can extract and summarize information related to the topic in different aspects based on the type of the topic.

Selected Publications and Outputs

1. Jing Jiang and ChengXiang Zhai, "A Two-Stage Approach to Domain Adaptation for Statistical Classifiers." In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*, pages 401-410, 2007.
2. Jing Jiang and ChengXiang Zhai, "Instance Weighting for Domain Adaptation in NLP." In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 264-271, 2007.
3. Jing Gao, Wei Fan, Jing Jiang and Jiawei Han, "Knowledge Transfer via Multiple Model Local Structure Mapping." In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'08)*, pages 283-291, 2008.
4. Jing Jiang and ChengXiang Zhai, "Exploiting Domain Structure for Named Entity Recognition." In *Proceedings of the Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'06)*, pages 74-81, 2006.
5. Jing Jiang and ChengXiang Zhai, "A Systematic Exploration of the Feature Space for Relation Extraction." In *Proceedings of the Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'07)*, pages 113-120, 2007.