Breaking the Barrier of Internet Information Acquisition

Question Answering Systems for Smartphone

By: Xiaoyan Zhu, Ming Li, Yu Hao

Final technical report
Submitted: 2014-08-15

Project Information
IDRC Project Number: 104519-006 / 104519-014
IDRC Project Title: IDRC Research Chair in Information Technology / International Research Chair Initiative
Country/Region: People's Republic of China / Canada

Full Name of Research Institutions:
Tsinghua University, People's Republic of China
University of Waterloo, Canada

Name(s) of Researcher/Members of Research Team: Need contact information
Xiaoyan Zhu, Department of Computer Science and Technology, Tsinghua University, P.R.China
Ming Li, David R. Cheriton School of Computer Science, University of Waterloo, Canada
Charles L.A. Clark, David R. Cheriton School of Computer Science, University of Waterloo, Canada
Weihong Li, Chinese Association of the Visually Impaired, P.R.China
Minlie Huang, Department of Computer Science and Technology, Tsinghua University, P.R.China
Bin Ma, Department of Computer Science and Technology, Tsinghua University, P.R.China
Mingxing Xu, Department of Computer Science and Technology, Tsinghua University, P.R.China
Yu Hao: Department of Computer Science and Technology, Tsinghua University, P.R. China

Abstract:
This project aims to overcome language and technology barriers to acquiring information through the internet. We intend to invent new techniques to simplify internet search processes by developing a natural language search engine, technology to facilitate cross language search (Chinese and English), and technology that would enable elementary mobile devices to search the internet.

Keywords:
Information technology, natural language-based search engine, question answering system, search engine for the blind.

ii. Research Problem

Information exchange is a key to social progress. Today, the most efficient method for information acquisition is through internet search engines. In China, (a) 1.5 billion people cannot read 19.2 billion English webpages. (b) While only 163 million people in China use the internet, 580 million Chinese people use cell phones. (c) There are 16 million visually impaired people in China who cannot use the traditional search engines conveniently even if they could read English. To solve problem (a), commercial search engines translate complete webpages which often results unreadable pages. An alternative solution is for the search engine to actually find the "answers" and translate the short answers only. For problem (b), can we allow 580 million Chinese cell phone users to search the internet without actually being on the internet? We notice a very unique situation in China: 528 million (out of 580M) cell phone users are very skilled short message users (iResearch.com report). Chinese text is much shorter than English, and short messages fit most basic cell phone screen perfectly. This provides an opportunity for a search engine to give short answers. That is: a cell phone user sends a query via a short message to our search engine, and the search engine finds the answer and sends it back as a short message. For problem (c), the issue is to be able to obtain a short answer, translatable to Braille or sound, so that the visually impaired can access the internet easily.

The key to solving (a)-(c) is a natural language search engine that gives concise answers, eliminating the multiple interactive processes in the current search engines. This research will address these inequities via an innovative "natural language search engine". The team will develop this next generation search engine technology based on a novel information distance theory, as well as their prior work on Braille systems and Question and Answer systems (QA).

As the research program evolved, we noted the rapid development and use of smart phones, which motivated a change in approach to (b). Initially we assumed that the demand for smart phones would be low due to cost but this assumption turned out to be false. The Chinese smartphone market doubled from 2011 to 2012 (from 24M to 44M smartphones sold in Q2), again doubled from 2012 to 2013 (Q2 from 44M to 88M smartphone sold) and still grew rapidly in 2014 (Q2 from 88M to 103M). Many young people in China use smartphones, which are generally cheap but fully functional. We accordingly adapted our commercialization model from short message QA to smartphone voice activated QA. However, the natural language issues behind the two applications stayed exactly the same.

iii. Objectives

Our four main objectives are outlined below with a brief summary of results achieved.

Objective 1: Training a total of 20 grad students in computer science in the specific areas of data mining, search engines, natural language processing and search algorithms at Tsinghua University and at the University of Waterloo, over the 5 year period.

We exceeded the objective. In the five years, 57 students were trained at Tsinghua University, 41 of whom graduated before this term. Ten students were trained at the University

of Waterloo.

Objective 2: Formalize and develop a Canada-China centre of excellence of internet information acquisition at Tsinghua University.

This centre was founded at Tsinghua University in 2009. In that year, we further developed the centre. In November 2010, we organized a QA-oriented workshop, where seven experts in the area gave talks. We also have held two seminars in our new centre. In Oct. 2011, we organized a Human-Robot Interaction workshop, where 18 experts from the field gave talks.

We have already published 40 papers on the project at top conferences (32) and top journals (8), including 6 co-authored papers.

Objective 3: Perform research on factoid query problems, complex and list query problems, query analysis, solution (answer) analysis, and a new indexing data structure that is suitable for question answering.

In 2010, at the University of Waterloo, we were building an index and passage retrieval system for QA. At Tsinghua, we worked on query classification, processing, analysis and post-processing and also developed a prototype system. In 2011, the system was upgraded from the baseline platform built last year. At the University of Waterloo, we collected data of English Community Question and Answering (cQA) from websites such as WikiAnswers and YahooAnswers, and built indexes. At Tsinghua, we carried out further research on the question analysis, answer evaluation and the integration of multiple sources. We also developed algorithms of question semantic analysis, concept relatedness analysis, interactive question recommendation, semantic passage retrieval from Wikipedia and product review mining and implemented our QAnswer platform.

During 2012, we upgraded the system from the platform developed the previous year and we broadened the system by incorporating knowledge as the key factor for information acquisition. At the University of Waterloo, we built English knowledge bases and translated Chinese questions into English in order to provide answers to Chinese questions through English knowledge. At Tsinghua, we carried out further research on information analysis, knowledge base construction, question mapping, and interactive interface. Algorithms of question paraphrasing semantic analysis, missing semantic conditions modeling, domain ontology and knowledge construction, question to knowledge mapping, dialogue management and finer granular entity review summary were developed and integrated into our QAnswer platform.

In 2013 and 2014, focusing on the research of domain knowledge construction and utilization as well as on domain-related semantic analysis, we developed a universal prototype framework of QA system for a specific domain. This means that, given domain knowledge and a domain-related text corpus, a QA system can be developed quickly and easily. This is a significant contribution to our goal of implementing our QA technique to various domains and applications. We co-operated with Tencent and the Union Medical College to develop applicable QA systems in the domain of entertainment and medical education. We also developed the complete RSVP QA system for smartphones.

Objective 4: Design a natural language question answering system QUANTA.

We built multiple systems which integrates answers from heterogeneous sources, such as a web page, cQA and Wikipedia, and provides customized answers for users in different scenarios.

iv. Methodology

In the past 5 years, we carried out research on how to assist users acquire information on the internet. For this ultimate goal, considering the various requirements of different application scenarios, we developed various algorithms and methods to solve problems pertaining to information measurement, semantic understanding, knowledge organizing and dialogue management. Most of the techniques shared common ground but differed in order to suit different purposes and to adapt to the continual and rapid changes of the Internet as well as changes of our view of the Internet information acquisition over the course of the 5-years period. The techniques were to be considered as a whole, as no single algorithm or method could be the ultimate solution.

We began by building a question answering platform to deal with the factoid questions. The platform was quite robust and attained a performance of 80% in the F-Measure, making it a state-of- the-art-system. We continually developed and upgraded the system in keeping with our research on knowledge management. However, this factoid system was quite limited in its capability of answering complex questions, which is common to most of the information acquisition scenarios. Then, we developed techniques to find the direct answer for non-factoid and domain independent queries, and built the QAnswer platform, integrating heterogeneous sources of information, including the user profile to provide the best answer. The platform was quite applicable and deployed in many real systems, such as the Flytech's Yudian. The take-off of smart devices spurred great needs in the information acquisition within vertical domains, which then required higher performance in precision and coverage. On the one hand, the domain applications accumulated huge volumes of data yet to be fully utilized, and on the other hand, there were fruitful research achievements on knowledge representation, acquirement and construction. In fact the vertical QAs for specific domains featured by high accuracy are badly needed by enterprises nowadays, whereas the construction periods are long and the costs are high, so that only big companies would have the capability to build such systems. For most enterprises, a customized QA system is still out of reach. However, our research in intelligent knowledge-driven dialogue systems has resulted in significant achievements in unified platform and systems for metrological, music, and medical applications.

In order to summarize our achievements, we present, in the following, our research work with reference to the theory of information distance metric, knowledge acquisition and utilization, semantic understanding, and the technology QA platforms.

1. Research on information distance metrics

1.1 Extensive research on information distance

Information distance is the universal measure between any two information-carrying objects [Li and Vitanyi: An introduction to Kolmogorov Complexity and its application, Springer,

2008 3rd Ed]. Based on this measure, we developed conditional information distances and information distance among multiple objects, and applied them to the text processing domain [1, 2, 9]. With the power of information distance, the similarity between the question and answer was estimated and the most appropriate answer was selected from the candidates [5, 10]. We also developed methods to generate answers from an information distance based summarization algorithm to obtain the summary of multiple candidates where typical and comprehensive sentences were both considered [6, 12, 13, 14]. Information distance was one of the basic theories that were helpful in information acquisition and may have direct effect on the measurement of distance between the requirement and the target.

The research in this direction yielded some important findings [5,9,32]. For example, our paper [5] was published in the *Communication of the ACM* in 2013. Papers [9,32] received best paper awards from COLING 2010 and ACL 2012 respective. *Communication of the ACM* is the top journal in computer science while COLING and ACL choose one best paper from over a thousand submissions. These papers have demonstrated the applications of information distance to computational linguistics.

1.2 Research of the semantic distance between concepts

Measuring the semantic similarity or relatedness of the conceptions was very important to understand the semantics of both queries and targeted answers. The intent of the question was usually implied by the focus words and their relationships. We proposed a novel method RCRank to jointly compute concept-concept relatedness and concept-category relatedness based on the assumption that information carried in concept-concept links and concept-category links could mutually reinforce each other [2]. Different from previous work, RCRank can not only find semantically related concepts but also interpret their relations by categories. This research involved creative work in information metrics based on the semi-structured encyclopedia corpus like wikipedia, which was open, popular and objective, contributed by internet users all over the world.

2. Knowledge Generation and Management

Knowledge is another important basis for information acquisition and intelligent systems. It is a special kind of information that is well structured and machine-understandable. Its representation, generation and utilization have been studied for decades since the AI developed. Up until now, in the era of Web 2.0 and big data, knowledge became more and more important and practical. Google developed the Knowledge Graph on the basis of Freebase, Yago, NELL, Conception Net and many pre-existing knowledge bases, which are still developing and which affect each other. We benefited from all the above works and focused our research on the Chinese knowledge base construction and its applications to information acquisition.

2.1 Knowledge gathering and ontology generation

User generated contents (UGCs) contain a large volume of high quality information. However, the information overload and diversity of different UGC sources limit their potential uses. We proposed a framework to organize information from multiple UGC sources using a topic hierarchy that is automatically generated and updated from the UGCs [34]. We explored

the unique characteristics of different kinds of UGCs, including blogs, cQAs, tweets, etc., and introduced a novel scheme to combine them. We also proposed a graph-based method to enable incremental updates for the generated topic hierarchy. Using the hierarchy, users can easily get a comprehensive, in-depth and up-to-date picture of their topics of interests. The experiment results demonstrate how information from multiple heterogeneous sources improves the resulting topic hierarchies. It also shows that the proposed method achieves better F-1 score performance in hierarchy generation compared to the state-of-the-art methods.

2.2 Chinese knowledge base construction

There are many state-of-art open knowledge bases, such as Freebase, Yago and Conception Net. However, as these are mostly in English, they lack information that is uniquely Chinese and therefore essential for a Chinese information acquisition system. We aimed to build a Chinese database that covers common knowledge, and knowledge of particular interest to the Chinese. We began by building a fundamental, high-quality knowledge base containing one million entities by incorporating data from Freebase, Wikipedia and Baidu Baike. Then we tried to infer patterns that would generate the labeled knowledge by coupling different extractors and constrains. As with the NELL project in Carnegie Mellon University (CMU), we employed new patterns to glean new knowledge from the web pages or User Generated Contents (UGC) such as blogs, tweets and cQA corpus. Through a semi-supervised bootstrapping method, we gathered world knowledge with little human interference. We implemented the methods to Baidu Baike and Baidu Zhidao (one of the most famous cQAs in Chinese), which generated about 0.5 million entries that were sampled and checked manually, the latter showing near- 80% accuracy.

2.3 Construction and application of the domain specific knowledge base

Knowledge is important for the understanding of the semantics of Natural Languages. For the state-of-art study, people have built many large scale knowledge bases such as CYC, Yago, DBPedia and Freebase. Those knowledge bases are horizontal and cover many domains, yet for every specific domain, the knowledge is not rich enough for practical vertical search engines. Therefore, in addition to the above Chinese knowledge base that covered common knowledge, we tried to gather domain specific knowledge. That knowledge was more detailed and unique for the featured application scenarios. We began by building, a fundamental high-quality knowledge base of 1 million entities by incorporating data from Freebase, Wikipedia and Baidu Baike. Then, some predefined databases composed by domain experts were adopted as supplements. We also tried to acquire patterns that would generate labeled knowledge by coupling different extractors and constraints from the UGCs. We then proposed a general framework to organize information from multiple UGC sources (blogs, tweets, and cQAs) using a topic hierarchy that was automatically generated and updated from the UGCs. A graph-based method was created and implemented to enable the incremental updating of the topic hierarchy to dynamically improve the semantic relations [34]. In this manner, we built multiple knowledge bases, including weather, entertainment, products, the mobile company business process and, most importantly, medical knowledge base.

3. Understanding the semantics

For the internet information acquisition task, semantic understanding remains a major problem. Since it is a mission impossible for computers to grasp the real semantics of the entire context, semantic understanding is approached practically and technically from different levels. A query's semantic may be represented by the document set that is similar to it, similar to how a traditional search engine would represent it. However, we wanted a natural language question to generate a precise answer, a challenge that required careful analysis of the query and the context of the target documents. We developed methods of semantic understanding of various levels ranging from question classification, paraphrasing, focus spotting and passage retrieval to knowledge graph mapping. These techniques ensured that the question semantics were understood well enough to retrieve the desired answer.

3.1 Question classification

This work aimed identify the type of the desired answer. Unlike content related system, such as the UIUC taxonomy, ours classified questions according to the user's purpose. We developed a boost pattern-based algorithm for the coarse classification and Markov Logical Network-based algorithm for the detailed classification. Compared with previous approaches, the proposed method combined the features of both soft patterns and the statistical n-grams, achieving better performance [19].

3.2 Sentence similarity measures and paraphrase

From a practical point of view, semantic understanding may be approached by determining that two sentences are identical semantically. Learning how to rewrite sentences is a fundamental task in natural language processing and information retrieval. We proposed a new class of kernel functions, referred to as string re-writing kernel, to address the problem. A string re-writing kernel measured the similarity between two pairs of strings, each pair representing re-writing of a string. It could capture the lexical and structural similarity between two pairs of sentences without the need of constructing syntactic trees. We further introduced an instance of string rewriting kernel which could be computed efficiently. Experimental results showed that our method could achieve better results on paraphrase identification and recognizing textual entailment, and was applied to our QA system to handle flexible natural language input [32].

3.3 Analysis of the missing semantic conditions

Users often omit some background information in their queries. With the incomplete or unclear query, search engines may return various irrelevant documents, which is unacceptable for our information acquisition system. Therefore, we focused our research on the analysis of certain important dynamic conditions, such as time and location. Taking the temporal factor as an example, we proposed a novel approach for the representation of temporal information, namely one in which temporal sensitivity is estimated from both word and context inexplicitly under a unified probabilistic paradigm. At the same time, the temporal scale is also analyzed in finer granularity. As an application, experiments on the question of retrieval demonstrate that time-sensitive queries are precisely detected and question-ranking performance is

improved effectively. This research received the Google Research Awards in 2012, the only award in China by Google.

3.4 Knowledge mapping of questions

Assuming that the knowledge base is a huge graph where concepts are nodes and relations are edges, we tried to map the question to a sub-graph of the knowledge base, which is a typical approach for semantic understanding Finding out the entities and their relationships in the question is considered semantically understandable by the machine, which may give out the answer directly through inference and database search or utilize the local knowledge to find the result from semi-structured data. Our work focused on the question and mapped the question to the entities and relationships in knowledge bases. For example, the question-understanding task to interpret a natural language question goes from (e.g., ”what country did Bruce Lee come from”) to three questions (?x, Bruce Lee, birthplace). The basic idea was to generate semantic contexts for the questions from large web databases, e.g., UGCs such as Freebase and community QA, to enrich and specify their meanings. A statistical author-topic model was used to describe the relationship between entities based on their contexts.

4. Answer generation

With the understanding of question semantics, the answer was retrieved or generated from multiple sources. Techniques were developed to select, verify, integrate and recommend the answers, which were also critical for the information acquisition system [17]. The methods described below were carried out based on the results of semantic understanding, and combined with each other to present a satisfactory answer.

4.1 Passage retrieval from semi-structured web pages

This method was designed to locate the most appropriate part of a document as the answer. Usually, there existed articles of high quality to answer a question, however, these were somehow too long to be the exact answer. With the help of knowledge mapping of the question, we extracted the specific passages from the related webpage to be the exact answer. In our method, the entities and their context information were leveraged to estimate the question’s topic distribution in Latent Dirichlet Allocation (LDA), and correspondingly found the most similar topics from the candidate passages segmented beforehand also by topics generated from the allocation [24]

4.2 Answer evaluation

cQA was one of the most important sources to get the answer. However, there is considerable ‘noise’ in this type of user-generated content. Therefore, the candidate answers from cQA needed to be evaluated before being presented to users. We proposed a learning-to-rank method based on the semantic analysis result. Four kinds of features were extracted to train the model: user-related features, semantic features, evidences from other related question-answer pairs and more importantly, the temporal features. Experiments showed that the refined features took account as many factors as possible in describing the answer quality and outperform traditional user-related features (credit scores on the page) [17].

4.3 Heterogeneous knowledge integration

Questions of different types are processed by multiple heterogeneous engines (such as web search, CQA, Wikipedia, and vertical search engines), and the results were selected or integrated according to the question semantics, source credibility, and answer quality, where a Bayesian paradigm of learning-to-rank was implemented to select the most appropriate answers then an information distance base summation was carried out to generate the final answer [17].

4.4 Answer recommendation and suggestion

The information acquisition task is actually a natural interaction process rather than question and answer. User may need help when the answers were not satisfactory. We have proposed two approaches to facilitate users in finding the best questions for both the user and the system.

The first approach proposed was Keywords to Questions (K2Q), assisting users to articulate and refine questions. Firstly, candidate questions and refinement words were generated from the set of input keywords. Secondly, after specifying some initial keywords, a user received a list of candidate questions as well as a list of refinement words. He or she could either select a satisfactory question, or select a refinement word to generate a new list of candidate questions and refinement words. We proposed a User Inquiry Intent (UII) model to describe the joint generation process of keywords and questions for ranking questions, suggesting refinement words, and generating questions that may not have previously appeared. An empirical study showed UII to be useful and effective for the K2Q task [28].

The second approach tried to generate next step suggestions when one turn of question and answering was over. The prompt was generated online according to the user question semantics and the knowledge related to the user's intent. A semantic clustering method was proposed to predicate the missing important information of user's questions, and prompted the user by rhetorical question. Improved from the traditional semantic clustering method in WordNet, we integrated the RCRank similarity measure estimated from Wikipedia and Baidu Baike to enhance the description capability and language fluency [17, 23].

4.4 Knowledge driven dialogue management

For information acquisition of the vertical domain, dialogue is an important interactive way to help users present their query intents along with the query process. Based on the previous approach of semantic understanding, the query as well as the user profile was mapped to a sub-graph of the entire knowledge graph. We proposed a knowledge-driven strategy to find the most effective path to the final answer, and accordingly provided the prompts or rhetorical questions for users to answer. The sub-knowledge graph was dynamically maintained as the dialogue process went on. Different from the traditional target-driven approach, our system was effective in considering all the related factors in a united information gain optimization paradigm, which was suitable to vertical information acquisition and easy to deploy.

5. Platforms and Applications

Based on all the above techniques, we have built information acquisition platforms that were domain independent/dependent according to different applications.

5.1 Domain independent information acquisition

In 2010, we built a prototype system QUANTA allowing people to ask factoid questions in natural language and QUANTA returned the information directly. It consists of multiple modules featured by indexing and passage retrieval system. We have designed a new indexing system for QA purposes only and a passage retrieval method for retrieving proper paragraphs for the QA practice, namely by using a new conditional random field model, and efficient algorithms.

Starting in 2011, we then developed QAnswer (http://www.qanswers.net:8080/cqa), which was in Chinese and domain-independent, which means that the search engine searched the web for the best answer. Integrating techniques in distance measure, question analysis, answer generation and answer integration, QAnswer utilized multiple sources on the internet and performed well in answering Chinese questions of all types. The system was deployed on the web and IM applications and provided API services to embed into third party products. For example, the Flytek Corporation integrated our result for its product YuDian, a personal mobile assistant.

5.2 General framework for domain-specific QA and its applications

With the rapid development of mobile devices, domain-specific QA systems were more focused than general information acquisition systems, such as SIRI, Rhino, and YuDian in China. There was a huge requirement of domain-related QAs from government, organizations and companies. Based on the techniques of knowledge generation, semantic analysis and answer generation, we proposed a general framework for domain-specific QA characterized by knowledge driven, pipelined data procession and fast deployment, which would help users to rapidly build their customized QA in no time with little human efforts.

As the implementations of the framework, we developed several practical vertical QAs, such as Weather, Vegetable Price, Mobile Map, Music, and so on. The Vegetable price applications, for example, allows farmers, retailers and consumers to identify vegetable prices in markets all over China. This information is now easily accessible and empowers both sellers and buyers to make informed decisions. For our cooperation with the Union College Research Center, we built a prototype medical QA that adopted natural language questions and generated answers regarding diseases, symptoms and medicines. Given the applications our research has open up, we received a two year extension of this project from IDRC to further our research and develop an auxiliary medical system for doctors in rural areas.

Additional funding

In addition to funds from IDRC and the Canada Research Chairs program, the project team was successful in securing the following grants that expanded the research program.

Project title	Funding Agency	Role in the project	Start date (year) / end date (year)	Value (indicate currency)

Research of Key Technologies in Intelligent Internet Information Acquisition	Tsinghua University	PI	2010-2013	2,000,000 RMB
Graph-based Text Mining Theories and Methodologies	NSFC	PI	2010-2013	2,100,000 RMB
Predicting Missing Semantic Conditions in Natural Queries	Google	PI	2011-2012	20,000 USD
Incorporating Knowledge Graph with Hierarchically Organized Social Media	Google	PI	2014-2015	50,000 USD

The funding from Tsinghua University focused on the research of basic theories relating to information acquisition, including information distance metrics, semantic understanding, and knowledge engineering. This funding enabled us to work with many renowned professors in the fields of AI. Some of the outputs from the fund directly affected our project, e.g., tools for Chinese NLP, the knowledge base infrastructure and indexing techniques.

The funding from the National Science Foundation of China (NSFC) aimed at the theoretical research of text mining from the probabilistic graph point of view. In this project, researchers tried to incorporate prior knowledge into the probabilistic graph model of words, entities, topics and documents. This model was applied to the tasks of entity linking, semantic relation extraction, sentiment analysis and recommendation, which were used for semantic understanding, knowledge construction and answer generation in our project.

The Google award "Predicting Missing Semantic Conditions in Natural Queries" aimed directly at the analysis of the missing conditions (e.g., temporal and location information) in a query, which was one of the research focuses in the field of information acquisition, and the outcome was directly used for the question understanding part of our project.

The Google award "Incorporating Knowledge Graph with Hierarchically Organized Social Media" proposed to combine the knowledge base with the huge volumes of UGCs to enhance the usability of the knowledge base. This ongoing project benefited our research in knowledge base construction and the semantic understanding of vertical domains, especially the medical domain.

Project outputs

List of trainees

	Trainees supervised by IDRC RC	Trainees supervised by the CRC	Trainees co-supervised by IDRC RC & CRC	Trainees supervised by other collaborators
Undergraduate	Fan WU, Cray, Shuyang LIN, Yilei YANG, Xin WU, Junpeng QIU, Weipeng HE, Yuxuan XIE, Xing SHI, Sicong ZHANG, Wenlong TU, Haoyu WANG, Xing XU, Peng JIANG, Zhiqiang GU			Chenguang WANG, FangziWANG, Shali LIU, Tan ZHANG, Yi YANG, Mingxing ZHANG, Xiaokai WEI, Wenjie YANG, Boxuan GUAN, Xi WEN

Masters	Yang TANG, Jingchen LIU, Chao HAN, Yangpeng LI , Mattia, Vincent, Hongwei JIN, Haoxiong TAO, Tanche LI, Nicolas, Yipeng JIANG, Yicheng LIU, Biao LIU		Kun Xiong, Di Wang, Borui Ye, Junnan Chen, Yahui Chen, Haocheng Qin, Wei Shao	
Doctoral	Chong LONG, Fangtao LI, Feng JIN, Fan BU, Zhicheng ZHENG, Lijing QIN, Po HU, Naitong YU, Xingwei ZHU, LI ZHAO, Jun FENG, Han XIAO, Lei FANG, Shouzhong TU, Daoyi LI, Yequan WANG		Chong LONG，Fan BU, Guangyu Feng, Xuefeng Cui, Xianglilan Zhang, Hongnan Wang	
Post-doctoral	Bin LIU, Cheng LING，Ting Wang		Anqi Cui, Yang Tang	

Throughout the program, we reported to IDRC on training environment we created through the program. On the provided scale, we consistently felt the IRCI significantly enhanced the training environment we could otherwise provide to your students.

	Not at All	Minimal	Moderate	Significant	Don't know
CRC				Yes	
IDRC RC				Yes	

Notable examples:

1. The IRCI financed the students to attend the international conferences and publish the papers on international journals.
2. CRC professor Ming Li undertook a great deal of informal student supervision in addition to his formal supervision.
3. The exchange of research experiences with the Canadian students helped a lot in algorithm research and system development.

Cumulative list of your research outputs by type

Type	Total number of research outputs
Journal Articles (published / accepted)	8
Journal articles (still in submission process)	4
Conference papers	32
Presentations (non-academic)	
Books	
Book chapters	
Newspapers / other media	

Theses	17
- MA / MSc	12
- PhD	9

Project Outcomes

During the execution of the project, Tsinghua University and the University of Waterloo exchanged a lot with each other through visits and emails. Together we built the QA system, published 6 co-authored papers (5 journal papers and 1 conference paper, which was selected the best paper among the 800 papers of that conference). The IDRC RC and the CRC jointly applied for an independent scientific research fund project of Tsinghua University, which has been executed well until now. We also organized a QA-oriented workshop in 2011, where seven top experts from the field were invited to give talks.

For the student training, 12 masters and 9 doctoral students have graduated. Our technology transfer goals have been furthered through employment. Among our graduates, 3 students now work for Google, 1 for Yahoo, 1 for Microsoft and 1 for IBM. Most of them pursuit the research topics related to this project, and clearly benefited from the knowledge and skills obtained from the research work here.

The research outputs consisted of the research, development and transference of the techniques about the intelligent information processing and QA platforms. During the development of the platforms, more than ten undergraduate students were taught basic algorithms of data mining and machine learning, trained for software engineering, experienced the process of real system development, and more than 20 bachelor's degree thesis also contributed. There were also about 30 master's and doctoral students who contributed to the algorithm research, from which more than 30 high quality papers were published, including one that won conference best paper and two conference best student paper. We have 2 awards from Googles, one is for the research of the missing information in queries, which is an important topic of semantic understanding; the other one is for incorporating the knowledge graph with hierarchically organized social media, which is about the application of the knowledge graph. We received these award because we made novel contributions on question semantic analysis and research of vertical domain QAs.

Furthermore, to apply our research, we built the open domain QA system: QAnswer (http://www.qanswers.net:8080/cqa) and vertical QA platform, based on which many vertical QA systems were developed, such as the weather forecasting, the navigation map, the vegetable price, the music and the medical QA. All systems were deployed to the Wechat platform for testing, and some techniques were transferred to Chinese companies such as Flytec and Tencent.

Some significant outcomes relating to our training, research and technology transfer goals are illustrated below:

Research & Technology transfer

- Developed applications of our natural language search engine with the China's Weather Bureau and China's Ministry of Agriculture.
- Created a start-up company called RSVP Technologies that attracted start-up capital

from the Chinese government as well as $2Million Canadian dollar VC funds. Now the company operates with offices in China and Canada. http://www.rsvptech.cn/about

Societal Impact

Collaborating with the Chinese Agriculture Ministry and the Central Weather Bureau, we have already released an Agriculture Price app that contains all agriculture product prices all over China. This is serving hundreds of million farmers in China.

While this product is for public interests only, our first commercial product (旅问旅答) will be online in April 2015. This will be a natural language QA system that provides tourism information for 3 billion tourists industry in China. This system is based on wechat. It answers any question from a tourist on road, from weather to travel to tour guide. We hope to reach 10 million users the first year it is on the market.

Future research – medical application

The continuation of this project will focus on medical information acquisition to serve rural doctors in low income areas. Based on our techniques of domain independent/dependent QA and knowledge base construction and application, we will build an applicable and reliable system to help doctors diagnose and treat illnesses. Our efforts will focus on the implementation of user intention understanding, construction of medical knowledge base (general medical information concerning the diseases, symptoms, medicines and therapies as well as the detailed regional epidemic information), limited inference of knowledge, intelligent answer generation, information pushing or recommendation and multimedia information indexing and retrieval (for medical images). We anticipate that rural doctors will benefit from the user friendly input interface either by natural language and disease images, accurate and authoritative answers or even suggestive solutions and carefully selected knowledge pushed for further study. This project will combine the power of academic research of information acquisition, medical experts from the Union Medical College and the local hygiene administration to set up a bridge for the inexperienced doctors to obtain the essential information and knowledge, thus finally benefit the local people in China.

ix. Bibliography

Research Output Bibliography

i. Journal Articles (published/accepted)

1. Long, C.; Huang, M. L.; Zhu, X. Y.; et al. 2010. A New Approach for Update Multi-document Summarization. Journal of Computer Science and Technology, Vol.24 No.4, pp739-749, 2010.
2. Bu, F.; Zhu, X.Y.; Li, M. 2011. A new multiword expression metric and its applications. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 26(1): 3-13 Jan. 2011. DOI 10.1007/s11390-011-1106-y
3. Long, C.; Zhang, J.; Huang M.L.; Zhu, X.Y.; Li, M.; Ma B. 2014. Estimating Feature Ratings through an Effective Review Selection Approach. Journal of Knowledge and Information Systems, 38(2): 419-446 (2014)
4. Jin, F; Huang, M.L.; Zhu, X.Y. 2011. Guided Structure-aware Review Summarization. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 26(4): 676-684 July. 2011. DOI 10.1007/s11390-011-1167
5. Tang, Y.; Wang, D.; Bai, J; Zhu, X.Y.; Li, M. 2013. Information distance from what I said to what it heard. Communications of the ACM, July 2013, 70-71.
6. Long, C.; Zhang, J.; Huang, M.L.; Zhu, X.Y.; Li, M.; Ma, B. 2014. Estimating feature ratings through an effective review selection approach. Knowledge Information Systems 38(2): 419-446 (2014)
7. Hu, P.; Huang, M.L., Zhu, X.Y. 2014. Journal of Computer Science and Technology, Vol.29, Num.3, May 2014
8. Hu, P; Huang, M.L., Zhu, X.Y. 2014. Patent Key Component Extraction with the Application of Patent Similarity Analysis. Journal of Computational Information Systems (2014) 5813 – 5820

ii. Journal Articles (submitted)

iii. Conference Papers

9. Bu, F.; Zhu, X.Y.; Li, M. 2010. Measuring the Non-compositionality of Multiword Expressions. In The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China. Best paper award, August, 2010.
10. Li, F.T.; Zheng, Z.C.; Bu, F.; Tang, Y.; Zhu, X.Y.; Huang, M.L. 2009. THU QUANTA at TAC 2009 KBP and RTE Track. Text Analysis Conference (TAC 2009), Gaithersburg, Maryland USA, November 2009.
11. Li, F.T.; Tang, Y.; Huang, M.L.; Zhu, X.Y. 2009. Answering Opinion Questions with Random Walks on Graphs. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL2009), Singapore. Aug. 2009.
12. Long, C.; Huang, M. L.; Zhu, X.Y. 2009. Multi-document Summarization by Information Distance. IEEE International Conference on Data Mining (ICDM), Miami, USA. 2009. pp. 866-871.

13. Long, C.; Zhang, J.; Huang, M.L. 2009. Specialized Review Selection for Feature Rating Estimation. IEEE/WIC/ACM International Conference on Web Intelligence (WI), Milan, Italy. 2009. pp. 214-221.
14. Long, C.; Huang, M.L.; Zhu, X.Y. 2009. Tsinghua University at the Summarization Track of TAC 2009. Text Analysis Conference (TAC), 2009.
15. Tang, Y.; Li, F.T.; Huang, M. L.; Zhu, X.Y. 2010. Summarizing Similar Questions for Chinese Community Question Answering P ortals. 2010 Second International Conference on Information Technology and Compu ter Science (ITCS2010), Ukrainian, July 24-25, 2010.
16. Zheng, Z.C.; Li, F.T.; Huang, M. L.; Zhu, X.Y. 2010. Learning to Link Entities with Knowledge Base. The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010), 2010.
17. Zheng, Z.C.; Tang, Y.; Long, C.; Bu, F.; Zhu, X.Y. 2010. Question Answering System Based on Community QA. LREC 2010 Workshop on Web Logs and Question Answering (WLQA 2010), Malta, 2010.
18. Li, F.T.; Huang, M.L.; Zhu, X.Y. 2010. Sentiment Analysis with Global Topics and Local Dependency. The Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010), Atlanta, Georgia, USA.
19. Bu, F.; Zhu, X.W.; Hao, Y.; Zhu, X.Y. 2010. Function-based Question Classification for General QA", EMNLP, 2010, MIT, Massachusetts, USA.
20. Li, F.T.; Han, C; Huang, M.L.; Zhu X.Y.; Xia, Y.J.; Zhang, S.; Hao,Y. 2010. Structure-Aware Review Mining and Summarization. The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.
21. Jin, F.; Huang, M.L.; Zhu, X.Y. 2010. A Comparative Study on Ranking and Selection Strategies for Multi-Document Summarization, CoLing 2010, August 23-27, Beijing, China.
22. Liu, J.C.; Huang, M.L.; Zhu, X.Y. 2010. Recognizing Biomedical Named Entities using Skip-chain Conditional Random Fields. ACL, workshop on Biomedical Natural language Processing, 2010, Uppsala, Sweden.
23. Bu, F.; Hao, Y; Zhu, X.Y. 2011. Semantic Relation Discovery with Wikipedia Structure. IJCAI 2011, July, Spain
24. Han, C.; Liu, Y.C.; Hao, Y.; Zhu, X.Y. 2011. Semantic Aspect Retrieval for Encyclopedia. CICLing 2011, February, Japan
25. Li, F.T.; Liu, N.; Jin, H.W.; Zhao, K.; Yang, Q; Zhu, X.Y. 2011. Incorporating Reviewer and Product Information for Review Rating Prediction. IJCAI 2011, July, Spain
26. Li, F.T.; Huang, M.L.; Yang. Y.; Zhu, X.Y., Learning to Identify Review Spam. IJCAI 2011, July, Spain.
27. Huang, M.L. Yang, Y; Zhu, X.Y. 2011. Quality-biased Ranking of Short Texts in Microblogging Services. IJCNLP, November, 2011, Thailand
28. Zheng, Z.C.; Si, X.C.; Chang, E.; Zhu, X.Y. 2011 K2Q: Generating Natural Language Questions from Keywords with User Refinement. IJCNLP, November, 2011, Thailand

29. Hu, P.; Huang, M.L.; Xu, P.; Li, W.; Usadi, A.K.; Zhu, X. Y. 2011. Generating Breakpoint-based Timeline Overview for News Topic Retrospection. ;In ICDM(2011)260-269
30. Huang, M.L.; Shi, X.; Feng, J.; Zhu, X.Y. 2012. Using First-order Logic to Compress Sentences. AAAI 2012, Toronto, Ontario, Canada.
31. Fang, L.; Huang, M.L. 2012. Fine Granular Aspect Analysis using Latent Structural Models (short paper). To appear In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-12). Jeju, Republic of Korea. July 8-14, 2012.
32. Bu, F.; Li, H.; Zhu, X.Y. 2012. String Re-writing Kernel. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-12). Jeju, Republic of Korea. July 8-14, 2012.
33. Li, F.T.; Pan, S.J.; Jin, O.; Yang, Q.; Zhu, X.Y. 2012. Cross-Domain Co-Extraction of Sentiment and Topic Lexicons. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-12). Jeju, Republic of Korea. July 8-14, 2012.
34. Zhu, X.W.; Ming, Z.Y.; Zhu, X.Y.; Chua, T.S. 2013. Topic hierarchy construction for the organization of multi-source user generated contents. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13). ACM, New York, NY, USA, 233-242.
35. Qin, L.J.; Zhu, X.Y. 2013. Promoting Diversity in Recommendation by Entropy Regularizer, IJCAI 2013, Beijing China. p 2698-2704
36. Fang, L.; Huang, M.L.; Zhu, X.Y. 2013. Exploring Weakly Supervised Latent Sentiment Explanations for Aspect-level Review Analysis. CIKM 2013.
37. Qin, L.J.; Zhu, X.Y. 2013. Functional Dirichlet Process. In Proceedings of the 22nd ACM international conference on Information and knowledge management. ACM, 2013
38. Qin, L.J.; Chen, S.Y.; Zhu, X.Y. 2014. Contextual Combinatorial Bandit and its Application on Diversified Online Recommendation. In Proceedings of SDM 2014, ; Best Student Paper runner up, Philadelphia, Pennsylvania, USA
39. Li, T.C.; Hao, Y.; Zhu, X.Y.; Zhang, X. 2014. A Chinese Question Answering System for Specific Domain. In Proceedings of WAIM 2014: 590-601, Macau, China
40. Feng, J.; Bian, J; Wang, T.F.; Chen, W.; Zhu, X.Y.; Liu, T.Y. 2014. Sampling Dilemma: Towards Effective Data Sampling for Click Prediction in Sponsored Search. In Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM2014), New York City, NY, US

iv. Theses

Ph.D

1. Chong Long, The Measurement of Information among Many Objects and its Applications, 2010, Tsinghua University
2. Fangtao Li, Research on Sentiment Analysis with Product Review, 2011, Tsinghua University

3. Feng Jin, Research on Document Summarization Algorithms and Their Applications, 2010, Tsinghua University
4. Fan Bu, Research on Information Metric for Internet Data, 2013, Tsinghua University
5. Zhicheng Zheng, Research on Ambiguity of User Queries, 2013, Tsinghua University
6. Lijing Qin, Statistical Models and Learning Algorithms for Recommender System, 2014, Tsinghua University
7. Po Hu, Research on the Key Technologies in Time-related Sequential Text Mining, 2010, Tsinghua University
8. Guangyu Feng, Approximating Semantics, to be completed 2015, University of Waterloo

Master

1. Yang Tang, Question Recommendation and Answer Summarization for cQA portals, 2010, Tsinghua University
2. Chao Han, Passage Retrieval and Question Adaptation Based on Semantic Method, 2011, Tsinghua University
3. Hongwei Jing, Research and Application on Sentiment Analysis with Product Reviews, 2012, Tsinghua University
4. Haoxiong Tao, Research and Implementation of Healthcare Question and Answering System for the Public, 2013, Tsinghua University
5. Tanche Li, A Chinese Question Answering System for Specific Domain
6. Vincent, Named Entity Disambiguation in Sentence, 2014, Tsinghua University
7. Di Wang, Learning automatic question answering from community data, 2012, University of Waterloo
8. A semantic distance of Natural language queries based on question answer pairs. 2014, University of Waterloo