Mobile Phone Text Processing and Question-Answering

Aisha Jilani
University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK

ABSTRACT

Mobile phone text messaging between mobile users and information services is a growing area of Information Systems. Users may require the service to provide an answer to queries, or may, in wiki-style, want to contribute to the service by texting in some information within the service's domain of discourse. Given the volume of such messaging it is essential to do the processing through an *automated* service. Further, in the case of repeated use of the service, the quality of such a response has the potential to benefit from a dynamic user profile that the service can build up from previous texts of the same user.

This project will investigate the potential for creating such intelligent mobile phone services and aims to produce a computational model to enable their efficient implementation. To make the project feasible, the scope of the automated service is considered to lie within a *limited domain* of, for example, information about entertainment within a specific town centre. The project will assume the existence of a model of objects within the domain of discourse, hence allowing the analysis of texts within the context of a user model and a domain model. Hence, the project will involve the subject areas of natural language processing, language engineering, machine learning, knowledge extraction, and ontological engineering.

Keywords: Natural Language Understanding, Text Message Processing

1: Prime Subject Area – Natural Language Processing

The prime subject area for this research paper is Natural Language Processing (NLP). NLP includes the core technical methods and theories that surround manipulation of human natural language using a computer machine. Roger Schank and Alex Kass suggest that in order to progress in NLP and design intelligent machine that understand natural language as human beings do, the researchers must, address such questions as how we (human beings) understand and represent the concepts that language can communicate, how we learn new concepts, and how we organize this knowledge in memory.

2: The Research Problem

The envisaged outcome of this (research) project is the theories behind designing a system that can understand a natural language message. This message will be received as a 'text message' that utilizes the Short Message Service (SMS) technology As computing becomes pervasive, use of SMS capability will not only prove handy but will also increase the usability of the system. Understanding the incoming 'text message' will involve application of Natural Language Understanding (NLU) techniques. Generating the reply 'text message' that is readable as natural language will benefit from application of technology in Natural Language Generation (NLG). The project is expected to deal with a question about an 'entity or something' within a closed-domain sent as 'text messages'. This can be classified into categories such as – definitions, where/what/who questions or temporal i.e. when questions. Furthermore it can also be open-ended question.

3: Keyword Searching Vs Question Answering

Today's search engines like Yahoo, Google, MSN have made information access lot easier. These perform key word search using techniques such as link analysis and counting number of words in the query and this has enhanced their capability of - *searching* information from most popular and lexically related pages. However, search engines are not designed to deal with natural-language questions (Roussinov et al, 2008). An emerging alternative to keyword-based web searching is automated question answering (Roussinov et al, 2006).

4: Question Answering Systems

Question answering (QA) has recently received attention from the information retrieval, information extraction, machine learning, and natural language processing communities (AAAI, 2002; ACL-ECL, 2002; Voorhees and Harman, 2000, 2001, cited by Brill et al, 2002). By acquiring AskJeeves.com (now called Ask.com) in July 2005, InterActiveCorp wanted to achieve a completely open-domain question answering system that (can) prove to be the holy grail of information access (Roussinov et al, 2008). General research has been going on in this area for long but Text REtrieval Conference (TREC) competitions have led to more specific research being carried out for the past few years. TREC has had a track dedicated for QA since 1999 (Anon, 2005) and has received a great deal of attention from the Computational Linguistics and Information Retrieval research communities in the last few years (e.g. TREC 2001–2003, cited by Soricut et al 2006). The definition of the task, however, is generally restricted to answering factoid questions (Soricut et al 2006) with a focus on open-domain systems. TREC reflects the information access paradigm shifting from keyword-based search to natural language driven navigation (Roussinov et al, 2006).

5: Restriction Imposed by SMS Technology

It is not possible to generate a paragraph-sized chunk of text as the answer under the limitations imposed by SMS technology itself. This is because the length of an SMS reply is limited to 160 characters only. This implies that the QA system under discussion should have the ability to generate an answer that is short, precise and still accurate. Complexity gets involved also because a 'text message' often uses slang and short words as compared to natural language. This makes it difficult to 'understand' a 'text message' as the language in use has no standard dictionary. Linguists agree that 'texting' effects individual's capability of using natural language. Grinter and Eldridge (2005) citing the work of Crystal (2001) discovered several mechanisms on how shortened form of words is derived by teenagers. They report use of traditional abbreviations or acronyms, ad-hoc shortened forms "made up" during the course of the communication, and the use of numbers and letters to represent sounds.

6.1: English Language – Challenging

Formal computer programming languages such as C++ or Java are strictly guided by the grammar rules and violation of simplest rule can generate an error. Spoken language also has set rules to be followed yet minor syntactic errors do not deteriorate the semantics of the communication. Basu (2008) argues that this flexibility of the natural language is what is difficult to achieve with language understanding computer systems because "What is communicated by natural language is not explicitly stated" (Schank et al, 2002). Natural Language researchers agree that at an ultimate level of natural language understanding, actually we need to understand both the syntax and the semantics of the sentence. Syntax is defined by the grammar of the language. If a sentence is syntactically well formed it helps in understanding the meaning of the sentence. But it is also important to note that sentences that are syntactically correct are not essentially correct in their meaning for example in the sentence *the dog ate my homework*; the sentence is grammatically correct but it is still meaningless. Basu (2008) believes that common sense knowledge, reasoning capability and experience with the use of language helps a human being interpret the meaning, hence – understanding a sentence.

6.2: English Language – Ambiguous

The inherent ambiguity in English language poses several problems for NLP systems. Some words in the language do not have a single meaning. To make sense of the meaning, the context, the word lies in needs to be understood e.g.

- *We gave the monkeys the bananas because **they** were hungry*
- *We gave the monkeys the bananas because **they** were over-ripen*

In the above sentence the use of word 'they' is ambiguous. The true meaning can be identified only if the properties of the objects – monkeys and bananas are known. English language does not clearly distinguish between the parts of speech in some sentences. For example in the phrase *'Pretty little girls' school'* it is hard to tell which word does the adjective – little apply to. This could apply to either the girls' or the school. It could be taken as to be connected to the word pretty as well. In other cases the tone or expression of words adds meaning to the overall sentence. For example in the sentence *'I never said she stole my money'* emphasis on each word changes the entire meaning of the sentence.

Technically there could be multiple parse trees for a single sentence which again needs deciding on which one to use. For instance for the sentence *'Fruit flies like a banana'*, two possible parse trees can be derived from the same grammar.

Consider the following sentence – *"Every respectable human worships a God"*. The given sentence is ambiguous. The message this sentence is conveying can be:

*Every respectable human worships a **(SINGLE)** God OR*
*Every respectable human worships his **(OWN)** God*

Due to ambiguity in the sentence it can be represented using first-order logic in at least two ways as follows:

For some X: for any Y: God(X)
AND IF (human(Y) AND respectable(Y)) THEN
worships(Y,X)

Another representation for the same sentence can be:
For any Y: (IF human(Y) AND respectable(Y) THEN
For some X: God(X) AND worships(Y,X))

It is also important to understand that the appearance and structure of the word changes in different situations. For example in plural form child becomes children whereas book becomes books. To achieve an ideal QA system such challenge posed by ambiguous nature of natural language need to be addressed.

7: Technical Review

The QA goal (of a typical NLP system) is to locate, extract, and provide specific answers to user questions expressed in natural language (Roussinov et al, 2008). Following is a brief introduction of a few systems designed in attempt to achieve the identified goals of natural language manipulation systems.

A recent QA system prototypes the model proposed by Dang et al. (2009). It allows users to search the library by using simple natural language queries. These questions relate to book's title, author, publisher and/or other (known) attributes of the eBooks. A series of steps (parsing, POS tagging, mapping) are performed on the user's query to understand it and then generate an appropriate answer.

AquaLog (2005) is an ontology-portable QA system for semantic web. It consists of a linguistic component that is used to parse and classify a natural language input query and represent it as a Query-Triple <subject, predicate, object>. These query-triples are generated by using GATE annotations for a given input query. AquaLog also associates additional features to the query triple using JAPE grammar. Different mechanisms are applied to resolve any ambiguity including using WordNet or, as a last resort, asking the user to select appropriate answer among different choices. User interaction could be a good option here but may not be suitable in some QA systems due to nature of communication e.g. mobile text messaging.

NSIR (2000 – 2005), pronounced "Answer", is a web-based question answering research prototype system designed by (CLAIR) group at University of Michigan. It is one of the early open domain QA system developed by an academic institute and it's under laying algorithms have been published in multiple publications. NSIR uses the existing web based search engines e.g. Yahoo, Google to retrieve the documents which are likely to contain the answer. It extracts entities by shallow parsing and returns small taxonomy of answer types. NSIR uses trainable classifiers and Brill's POS tagger to identify certain words in the sentences that helps determine the semantics of the question which in return helps in determining the semantics of the expected answer. Before answers are returned to NSIR users, they are ranked according to a set of machine learning techniques, including the proximity algorithm and probabilistic phrase ranking (Radev et al., 2002). Research experiments have shown NSIR capable to return exact answers or snippets of text containing the answer (Roussinov et al, 2008).

An article in an online magazine *Information Week*, dated 28th March 2005, had mentioned ASU QA (2002) system as one of the most promising directions in the "Search of Tomorrow" (Roussinov et al, 2006). ASU QA is a research prototype designed by Arizona State University. It uses trained regular expressions to identify the certain grammatical categories (e.g. "what is", "who is") of the user question. Instead of using typical tools ASU QA uses pattern-matching mechanism for named-entity-identification in the extracted answers. ASU QA was demonstrated in 2005 to help locate potentially malevolent online content, potentially helping law-enforcement (Roussinov et al, 2008). Alongside other research tools ASU QA has been used for several years for research purposes in $2 million project supported by NASA.

Falcon (2000) had been identified as one of the most successful systems presented in TREC9 (2000) track (Roussinov et al, 2006) that makes use of open domain linguistic resource WordNet. NSIR and ASU QA only use grammatical and semantic *types* whereas Falcon goes an extra leap; it classifies the identified words in the question in pre-built semantic *types* (like person, place, profession, date, etc.) and *sub type* (such as musician, writer, politician etc.) in an attempt to understand the semantics of the user question which in return helps in identifying the potential *correct* answer to the question. Falcon uses purpose built Information Retrieval (IR) techniques to find useful paragraphs of text within the documents.

QA LaSIE (2000) is IE system designed for research purpose by Sheffield University. It is an extension of LaSIE (Large Scale Information Extraction) system that was originally designed for participating in message understanding conference (MUC6 & MUC7). LaSIE did not have grammar rules to handle sentence posing questions this feature was added on hence the name QA LaSIE. This capability was achieved by introducing special semantic predicate module to identify entity under question.

Ask MSR is still a prototype, although Microsoft is trying to improve it and it may be launched commercially under the name AnswerBot (Anon, 2004). AskMSR uses the web as data repository and depends on data redundancy for generating answers to questions. The AskMSR team stresses on achieving harder things by simpler means. AskMSR generates several rewrite strings for the question assuming them to be sub-strings (sometimes by just ANDing the words appearing in the questions) of the expected *correct* answer. Question type determination has been widely used (Brill et al., 2001) question transformation and question expansion are also commonly used, via manually crafted question-to-query transformations (Brill et al., 2001).

START (1993) (SynTactic Analysis using Reversible Transformations) was developed by Boris Katz at MIT's Artificial Intelligence Laboratory (Anon, no date). START in contrast to other QA systems has the capability to bring up images whose annotations can be matched to the question. START learns from analysing information presented in English. This information is then stored in the Knowledge Base as if *digested summary* of the syntactic structure of the sentences learnt. The question sentence is analysed in a similar way and the results are matched with the information stored in the knowledge base. This matching is performed using sophisticated linguistic techniques and syntactic patterns hence achieving more than just keyword matching.

8: Conclusion

Technical review of different QA system models and methodologies reveals that a typical QA system consists of different components to accept a natural language question from a user and deliver its answer(s) back to the user. These components are used to *understand* the user question by passing it through successive NLP processes and then *represent* it in machine understandable format like parse trees or first order logic. This representation needs to be compatible with the knowledge representation so that it can be mapped to an answer using different mapping techniques available. This mapping can be probabilistic, deterministic or a combination of both.

A QA system is built on top of a knowledge base that is used to extract the potential answer(s) to the user question. This knowledge base is either built specifically for the system (domain-specific knowledge base for close domain systems) or an existing resource (such as World Wide Web (WWW) or text repository) can serve as a knowledge base (domain-independent knowledge base for open domain systems). Different tools and techniques are available to implement each component but their

selection is influenced by the nature of the system under design. A block representation of a typical QA system as understood is shown in the diagram below:

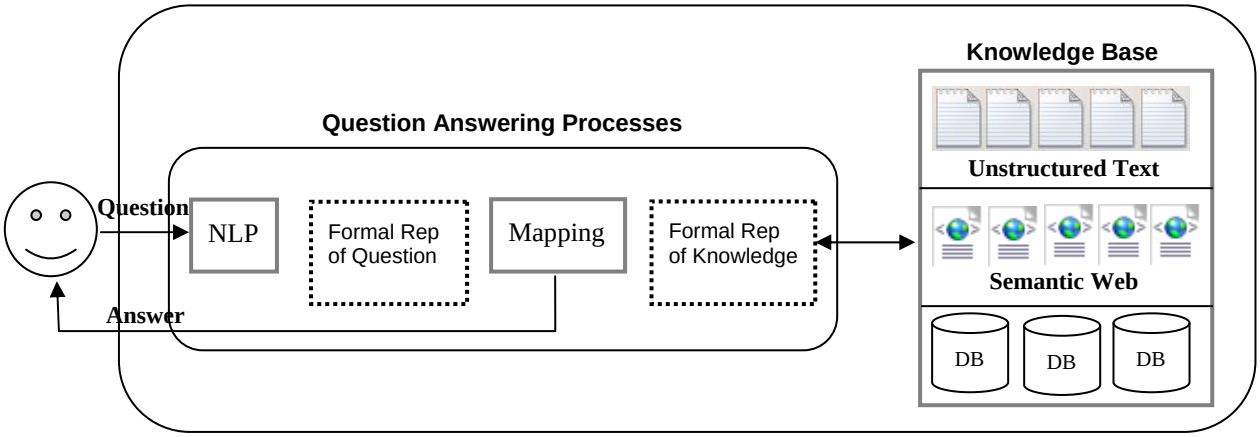

Figure 3: Block Diagram of a Typical Question Answering System

Early statistics show that, “Question Answering (QA) systems can answer nearly 85% of factoid questions” (Lin et al, 2003). Yet none of the above reviewed systems benchmarks QA system probably because each of them attempts to explore and experiment different aspects of QA.

Despite all up-to-date research in the subject area, researchers still agree that “parsing a sentence written in natural language is yet one of the difficult computing challenges and understanding the meaning of the sentence is proving even more difficult” (Gaševic et al, 2006). Having said this, it is also true that “several well-known futurists believe that computers will reach capabilities comparable to human reasoning and understanding of languages by the year 2020” (Roussinov et al, 2006).

THE END

Bibliography & References

Bibliography

1. Boose, J. and Gaines, B. (1989) Knowledge Acquisition for Knowledge-Based Systems: Notes on the State-of-the-Art; *Kluwer Academic Publishers, Boston -Machine Learning, 4, pp. 377-394* [online] Available from: Google Books < http://scholar.google.co.uk>

2. Liou, Y. (1992) Knowledge Acquisition: Issues, Techniques and Methodology, [online] Available from: ACM Digital Library < http://portal.acm.org/dl.cfm>

3. Scott, S. and Gaizauskas, R. (2001) QA-LaSIE: A Natural Language Question Answering System, *E. Stroulia and S. Matwin (Eds.):* AI 2001, LNAI 2056, pp. 172-182, [online] Available from: Springer Link <http://www.springerlink.com>

4. Greenwood, A. Roberts, I. and Gaizauskas, R. (2003) The University of Sheffield TREC 2002 Q&A System, *TREC 2002* [online] Available from: TREC Website <http://trec.nist.gov/pubs>

5. Lecturer. Ng, A (2008) Series of lectures on Machine Learning; Stanford University [online] Available from: Stanford University Channel on YouTube <http://www.youtube.com/stanford >

6. Lecturer, Basu, A. (2008) Lecture Series on Artificial Intelligence; *Department of Computer Science and Engineering, I.I.T, Kharagpur* [online] Available from: NPTEL <http://nptel.iitm.ac.in.>

References

1. Harabagiu S. et al (no date) Falcon: Boosting Knowledge of Answer Engines [online] Available from: <http://www.cse.unt.edu/~rada/papers/harabagiu.trec00.pdf>

2. Sapir, E (1921) Language: An Introduction to the Study of Speech; *Chapter 5. Form in Language: Grammatical Concepts* [online] Available from: <http://www.bartleby.com/186/5.html>

3. Simmons, R. (1969) Natural Language Question- Answering Systems: 1969 [online] Available from: ACM Digital Library **<** http://portal.acm.org/dl.cfm**>**

4. W. Chafe (1992) The Importance of Corpus Linguistics to Understanding the Nature of Language. In Jan Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, 79-97. Berlin: Mouton de Gruyter. [online] Available from: Google Books < http://books.google.co.uk>

5. Schank, R. And Kass, A. (2002) Natural Language Processing: What's Really Involved? [online] Available from: <http://www.aclweb.org/anthology-new/T/T87/T87-1023.pdf>

6. Katz, B. Lin, J and Quan, D. (2002) Natural Language Annotations for the Semantic Web; *Proceedings of the International Conference on Ontologies, Databases, and Application of Semantics (ODBASE 2002)*, October, 2002. [online] Available from: <http://groups.csail.mit.edu/infolab/publications/Katz-etal-ODBASE02.pdf>

7. Radev, D. Qi, H. Wu, H. Fan, W. (2002) Evaluating Web-based Question Answering Systems [online] Available from: <http://clair.si.umich.edu/~radev/papers/NSIR.pdf>

8. Brill, E. Dumais, S. and Banko, M. (2002) An Analysis of the AskMSR Question-Answering System, *Microsoft Research*, [online] Available from: <http://research.microsoft.com>

9. Chen, A. R. Diekema, M. D. Taffet, N. McCracken, N. E. Ozgencil, O. Yilmazel, E. D. Liddy (2001) Question answering: CNLP at the TREC-10 question answering track; *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)* – Cited by Brill, E. Dumais, S. and Banko, M. (2002) An Analysis of the AskMSR Question-Answering System, *Microsoft Research*, [online] Available from: <http://research.microsoft.com>

10. Brill, E. (2003) Processing Natural Language without Natural Language Processing; *Microsoft Research* [online] Available from: Springer Link <http://www.springerlink.com>

11. Grinter, R. And Eldridge, M. (2003) Wan2tlk?: Everyday Text Messaging; *Proceedings of the SIGCHI conference on Human factors in computing systems,* (2003) pp. 441 - 448 [online] Available from: ACM Digital Library **<** http://portal.acm.org/dl.cfm**>**

12. Soricut, R and Bril, E. (2006) Automatic question answering using the web: Beyond the Factoid; *Springer Science + Business Media Inc. 2006* [online] Available from: Springer Link <http://www.springerlink.com>

13. Roussinov, D. Fan, W. and Robles-Flores, J (2008) Beyond keywords: Automated question answering on the web, *Communications of the ACM*, September, 51 (9), pp. 61 – 65 [online] Available from: ACM Digital Library < http://portal.acm.org/dl.cfm>

14. Wintner, S. (2009) What Science Underlies Natural Language Engineering; Association for Computational Linguistics 2009 [online] Available from: <http://www.mitpressjournals.org>

15. Dang, N. And Tuyen, T. (2009) Natural Language Question Answering Model Applied To Document Retrieval System; *World Academy of Science, Engineering and Technology* (2009) 51 [online] Available from: <http://www.waset.org/journals/waset/v51/v51-7.pdf >