Proceedings of the First IEEE International Conference on
Semantic Computing (ICSC 2007), 439–445. ©2007 IEEE

Answering English Questions using Foreign-Language, Semi-Structured Sources

Boris Katz, Gary Borchardt, Sue Felshin, Yuan Shen and Gabriel Zaccak
MIT Computer Science and Artificial Intelligence Laboratory
{boris, borchardt, sfelshin, yks, gabi}@csail.mit.edu

Abstract

Despite continuing advances in machine translation technology, users who lack familiarity with particular foreign languages have no good way to find information in those languages. In this paper, we present a technical framework and implemented system that answers English questions on the basis of information in foreign-language, semi-structured sources such as websites. This work helps users locate, with high precision, relevant segments of foreign-language information, and then makes use of existing machine translation services to present that information in English. The resulting technology extends an approach embodied in the START information access system and its supporting Omnibase uniform data access system, and it has been applied to several Chinese and Arabic websites.

1. Introduction

The Web contains valuable information in many languages, and even though there are software tools that can form approximate translations of foreign-language text into English, users who cannot read particular languages and who are unfamiliar with particular foreign-language sites have no way to *find* the specific elements of information that are relevant to their needs. This is the case not only for conventional foreign-language Web pages, but it is also true for "deep Web" interfaces, where materials are accessed through form-based interactions.

We have developed a capability to answer English questions on the basis of foreign-language material in semi-structured websites. This work extends into a new arena an approach demonstrated by the START [1][2] and Omnibase [3] systems as used to answer English questions on the basis of English-language material in semi-structured sources.

At present, our foreign-language capability operates in connection with several Chinese and Arabic websites. Figure 1 illustrates the capability used to answer a question based on material in a Chinese-language site.

Figure 1. The START system answering a question by retrieving material from a Chinese-language website

Three key components of the strategy presented in this paper are: (1) the use of *natural language annotations*—content-describing phrases and sentences—to characterize information sources and the

questions they answer, (2) the use of *nested ternary expressions* as an underlying representation for natural language questions and assertions, and (3) the modeling of specific information sources using an *object–property–value* data model. These techniques have previously been applied only to English-language information sources, yet their underlying nature is language-independent, and thus it has been possible to apply these techniques to foreign-language information sources with relatively few changes. Indeed, we are not aware of other systems described in the literature that answer English questions on the basis of foreign-language, semi-structured sources, independent of the techniques employed.

The work described in this paper provides a complementary capability to that provided by cross-lingual question answering systems, as described, for example, in [4], [5] and [6]. These systems supply broad coverage in answering questions on the basis of information in large collections of unstructured text; however, given the breadth of coverage for these systems, it is difficult for them to provide high precision in question answering. The techniques described in this paper apply in a more focused context to selected, semi-structured datasets and provide high precision question answering over those datasets.

Sections 2 and 3 of the paper describe the three key strategies of the approach—natural language annotations, the ternary expression representation, and the object–property–value data model—and their use in the START and Omnibase systems. Section 4 then describes extensions to the approach required to answer English questions on the basis of information in foreign-language, semi-structured sources. Section 5 continues with a description of advanced question answering capabilities that have been applied in the context of foreign-language, semi-structured sources, and Sections 6 and 7 describe continuing research and contributions of the work.

2. The START system

START [1][2] is a publicly-accessible information access system that has been available for use on the Internet since 1993.[1] START answers natural language questions by presenting components of text and multi-media information drawn from a set of information sources that are accessed remotely through the Internet or hosted locally. These sources contain structured, semi-structured and unstructured information.

As originally configured during the initial stages of its development, START served to answer English questions on the basis of English statements that had been previously submitted to the system, and this operation underlies much of START's current capabilities as well. When START is presented with an English statement for processing, it parses the statement and encodes it in the form of a set of *nested ternary expressions* [1]. One can think of the resulting entry in START's knowledge base as a "digested summary" of the syntactic structure of the English sentence. User-submitted questions are then analyzed in the same manner and matched against stored assertions in the knowledge base. Matched assertions are then retrieved and expressed as English responses. Because matching occurs at the level of syntactic structures, linguistically sophisticated machinery such as synonymy, hyponymy, ontologies, and structural transformation rules can all be brought to bear on the matching process. Linguistic techniques allow the system to achieve capabilities far beyond simple keyword matching. In particular, the use of structural transformation rules, in both forward and backward modes, enables the system to find matches despite significant differences in expression that arise from alternate realizations of the arguments of verbs and other structures [7].

The START system bridges the gap between sentence-level text analysis capabilities and the full complexity of unrestricted natural language (and multimedia information) by employing *natural language annotations* [2]. Annotations are computer-analyzable, natural language sentences and phrases, typically composed by humans to describe the contents of individual information segments or sets of parallel information entities within information sources. START analyzes natural language annotations in the same fashion as any other sentences, but in addition to creating the required representational structures, the system also produces special pointers from these representational structures to the information segments or sets of segments summarized by the annotations. For example, an HTML fragment about clouds on Mars may be annotated with the following English sentences and phrases:

```
Clouds exist on Mars.
Martian clouds are composed of water and carbon
    dioxide.
...
```

START parses these annotations and stores the parsed structures (nested ternary expressions) with pointers back to the original information segment. To

[1] http://start.csail.mit.edu/

answer a question, the user query is compared against the annotations stored in the knowledge base. If a match is found between ternary expressions derived from annotations and those derived from the query, the corresponding annotated segment is returned to the user as the answer. For example, annotations like those above allow START to answer the following questions:

Are there clouds on Mars?
Do clouds exist on Mars?
What is the chemical composition of Martian clouds?
Do you know what clouds on Mars are made of?

Figure 2 presents an example of START answering such a question.

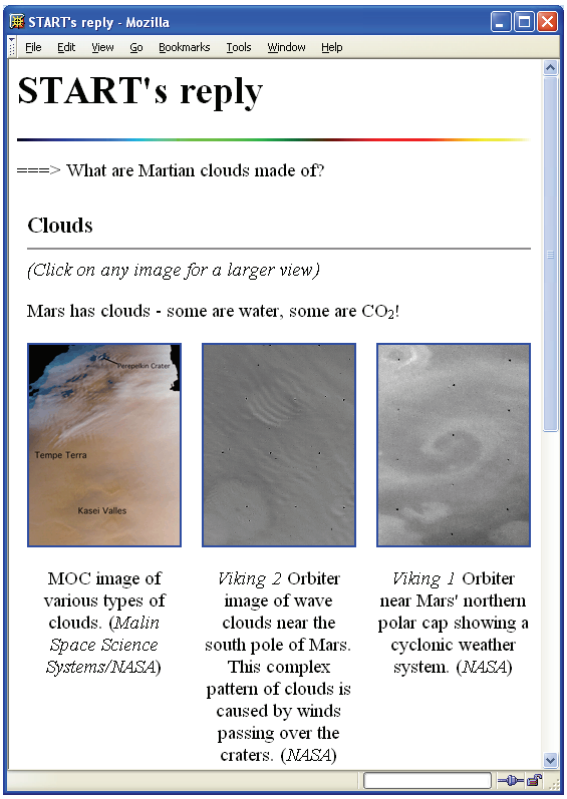

Figure 2. START answering a question using annotation-based matching

With large information sources, of course, it is impractical to manually annotate each item of content. However, sources of all types—structured, semi-structured and unstructured—can contain significant amounts of parallel material. *Parameterized annotations* address this situation by combining fixed language elements with "parameters" that specify variable portions of the annotation. As such, they can be used to describe whole classes of content while preserving the indexing power of non-parameterized annotations. As an example, the parameterized annotation (with parameters in italics)

number people live in *city*.

can describe, on the data side, a large table of population figures for various cities. On the question side, this annotation, supported by structural transformation rules, can recognize questions submitted in many forms:

How many people reside in Chicago?
Do many people live in Pittsburgh?
What number of people live in Seattle?
Are there many people living in the Boston area?

Additional parameterized annotations may be included that describe the population figures in other ways (for example, using the terms "population" or "populous"), and additional elements of the annotations may be parameterized. As a result, a large number of different questions can be answered using a small number of parameterized annotations. For example, with further parameterization, a single annotation can answer questions about area, population density, elevation, and other quantities in addition to population.

3. The Omnibase system

The Omnibase system [3] supports the START system by retrieving answers from a variety of semi-structured and structured sources in response to queries generated by START. Omnibase acts as an abstraction layer that provides a uniform interface to disparate Web knowledge sources. Omnibase models individual sources as collections of *objects*, with each object having one or more *properties* that have particular *values*. Parameterized annotations serve as the interface between START and Omnibase's object–property–value data model, allowing the combined systems to answer questions about a variety of topics such as a country's population, area, GDP or flag; a city's population, location or subway map; or a famous individual's place of birth, date of birth, or spouse. The object–property–value data model is more generally applicable than it may appear on the surface, as many object–property questions can be cast with diverse phrasing; e.g., "What is Angela Merkel's date of birth?" can be phrased as "When was Angela Merkel born?", "What is Argentina's size?" can be phrased as "How big is Argentina?", and so forth.

Figure 3 illustrates a question answered by START, utilizing support from Omnibase.

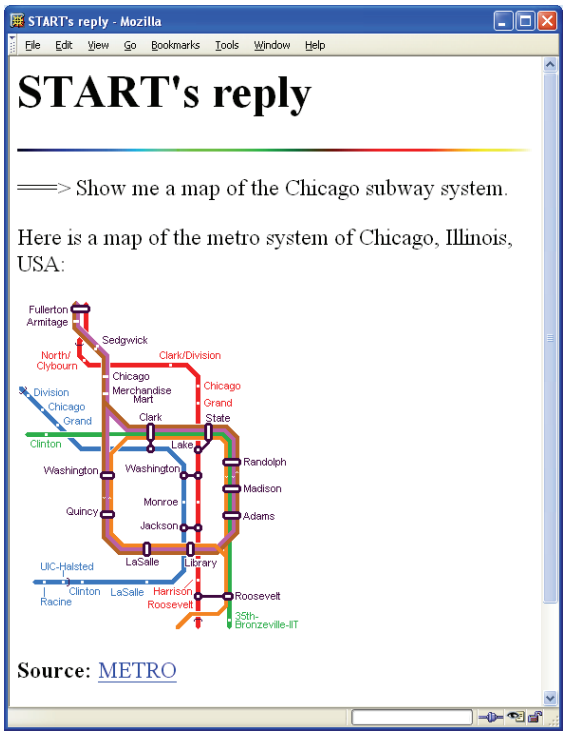

Figure 3. START and Omnibase answering a question using material from an English source

In order to successfully match input questions to parameterized annotations, START must know which terms can be associated with any given parameter. Omnibase supports this need by acting as an external gazetteer for source-specific terminology, with variants of terms being calculated automatically from objects' names, extracted from semi-structured material in information sources, or manually defined. This maintains the integrity of the abstraction layer: information source terminology is kept together with information source processing.

Omnibase's use of the object–property–value data model applies equally well to fixed, semi-structured websites and to "deep Web" sources that are accessed through special query languages or interactive form-based interfaces. When START transmits an object–property query to Omnibase, Omnibase executes an access script associated with the information source in question, and the access script may obtain individual elements of information directly from a static Web page, extract data from a local data source, or obtain dynamic information or otherwise "hidden" information by interacting with a query interface.

4. Modeling foreign-language sources

The realization that the object–property–value data model generalizes easily to foreign-language, semi-structured sources is a particularly interesting outcome of this research effort. As with English sources, application of the object–property–value data model involved the creation of a set of associated natural language annotations in START, plus the composition of a set of information access scripts for the properties covered by the model. In addition, application of the object–property–value model to foreign-language sources involved the creation of a mechanism by which START can process instances of foreign-language terms within English questions submitted to the system.

Foreign-language words can potentially be included within English questions in any of three ways. For example, the 1985 Japanese movie "Tampopo" can be represented

- in the original foreign language as タンポポ or 蒲公英,
- in transliteration as "Tampopo", or
- in translation as "Dandelion".

Our initial implementation of question answering from foreign-language sources supports foreign-language terms in transliteration. We extended the lexicon available to START and Omnibase so that it contains a collection of pertinent, transliterated foreign-language terms. A combination of automatic and manual techniques was employed to extract foreign-language terms from each targeted information source. These terms can be interpreted as instances of particular classes, according to the object–property–value data model. For example, one such class corresponds to regions in China, and for this class we enumerated instances such as 上海 (Shanghai) and 河南 (Henan) and entered their transliterations into START's lexicon.

A significant body of work exists in the area of machine transliteration, involving either automatic generation of transliterated words on the basis of spelling or sound patterns [8][9][10] or learned transliteration from name pairs and parallel corpora [11][12][13]. For Chinese–English name transliteration, we used the pinyin4j open-source transliteration package[2] to convert Chinese characters into tone free Pinyin, then applied a rule-based mechanism to handle special cases. For Arabic–

[2] http://pinyin4j.sourceforge.net/

English name transliteration, we used Google's machine translation engine[3] manually augmented where necessary.

With these elements in place, when the user enters a relevant request, START will recognize foreign-language terms within the request and use the classifications of those terms along with the ternary expression structure of the query to match an appropriate parameterized annotation and issue a suitable object–property query to Omnibase. In turn, Omnibase will execute an appropriate access script in order to retrieve a fragment of foreign-language text. The final step is to obtain an approximate English translation of the retrieved text fragment using machine translation software and present both the original fragment and its translation to the user. For machine translation, we are currently making use of translation engines offered by BBN [14] and by Google[3].

Figure 4 illustrates the end-to-end question-answering process in conjunction with the Arabic-language website *Rafed*.

Figure 4. START answering a question by retrieving material from an Arabic-language website

[3] http://translate.google.com/translate_t

5. Advanced question answering

In some cases, complex questions can be decomposed syntactically into simpler questions that can be answered independently. The ternary expression representation can be used as a guide in the determination of syntactically valid subquestions, and the results of processing for the subquestions can be fused to provide explanatory answers to the user. We have explored this kind of syntactic decomposition in the context of English-language resources [15], and this approach generalizes to the context of foreign-language resources as well.

Figure 5 illustrates START's handling of the complex question "How far is the capital of Henan from Seoul?" To answer this question, START first processes the subquestion "What is the capital of Henan?" The answer to this subquestion, 郑州 in the original Chinese, is transliterated as "Zhengzhou", and START proceeds by consulting its own knowledge base to determine the distance between Seoul and Zhengzhou.

Figure 5. START answering a complex question by retrieving material from a Chinese-language source and START's knowledge base

Separately, a number of other question answering mechanisms present in START can be generalized to use with foreign-language sources. These include

START's ability to process elided questions by borrowing terms and structures from previous questions, its ability to suggest corrections for misspelled words, its ability to utilize inexact matches between questions and annotations in the absence of exact matches, and its ability to generate responses with clickable links that contain pre-generated English "drill-down" questions for follow-on questioning by the user.

6. Continuing research

We are in the process of integrating additional Chinese and Arabic sources into our implementation, and we plan to include sources in other languages as well. In addition, we plan to extend our support for the specification of foreign terms by allowing users to enter English translations of those terms as well as entering the foreign-language terms in their native character sets.

Also, we are developing a new mechanism for processing syntactically-complex questions involving multiple information sources written in the same foreign language. This mechanism will bypass the unnecessary step of translating the intermediate results of the syntactic decomposition process into English. For example, if the answer to one subquestion is a Chinese term, and that term is needed to form an object–property query to a separate Chinese-language source, then the initial subquestion response can be retained in Chinese for use with the second source, thus avoiding the potential ambiguity that might arise from transliteration and subsequent reverse transliteration of the initial Chinese-language answer.

For each new semi-structured source to be integrated, it is necessary to compose a set of natural language annotations that describe the types of information contained within that source, as well as access scripts for object properties related to that source. Some parts of this process can be automated, and we have developed two systems, Hap-Shu [16] and more recently, Wrapster [17], to automate the creation of access scripts. We have applied these tools to English semi-structured sources and are planning to apply them to foreign-language semi-structured sources in the near future.

7. Contributions

Valuable foreign-language information exists in abundance on the Web and in other sources, yet users have no good way to find this information. This paper offers a partial solution to this problem in the form of a technical framework for answering English questions on the basis of information in semi-structured, foreign-language sources. Within this framework, natural language annotations are used to specify information source content succinctly, such that a question answering system can provide high-precision access to specific components of information within the sources. Matching of questions to natural language annotations is greatly facilitated by the use of nested ternary expressions as an underlying representation, as is the case when dealing with English-language sources. In addition, individual foreign-language resources can be modeled using an object–property–value data model, which facilitates tight integration with natural language annotations and allows for straightforward composition of access scripts for those sources.

A related contribution is the description of an initial implementation of the elaborated technical framework, involving application of the START information access system and the Omnibase uniform data access system. This implementation has demonstrated that it is possible to provide access to foreign-language semi-structured information in much the same way that a system can provide access to English semi-structured information, given the insertion of additional components for handling foreign language terms within questions and for generating approximate translations of passages selected for presentation as responses to the user.

The ability to answer English questions on the basis of foreign-language, semi-structured sources significantly extends the reach of the START system, allowing its users to access not only a much broader range of information, but as well information that has been articulated from vastly different cultural perspectives. This accomplishment is attributable to three key strategies that have been generalized from question answering over English-language semi-structured sources: the use of natural language annotations to match questions to information source content, the underlying representation of questions and assertions in terms of nested ternary expressions, and the characterization of information source content using the object–property–value data model.

8. Acknowledgements

This work is supported in part by the Disruptive Technology Office as part of the AQUAINT Phase 3 research program.

9. References

[1] B. Katz, "Using English for Indexing and Retrieving," in *Artificial Intelligence at MIT: Expanding Frontiers, v. 1*, Cambridge, Massachusetts, 1990.

[2] B. Katz, "Annotating the World Wide Web Using Natural Language", *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*, Montreal, Canada, 1997.

[3] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, A. J. McFarland, and B. Temelkuran, "Omnibase: Uniform Access to Heterogeneous Data for Question Answering", *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB '02)*, Stockholm, Sweden, 2002.

[4] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo and M. de Rijke, "The Multiple Language Question Answering Track at CLEF 2003," *Working Notes for the CLEF 2003 Workshop*, Trondheim, Norway, 2003.

[5] Y. Sasaki, H.-H. Chen, K.-h. Chen and C.-J. Lin, "Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1)," *Proceedings of the NTCIR-5 Workshop Meeting*, Tokyo, Japan, 2005.

[6] Y. Sasaki, C.-J. Lin, K.-h. Chen and H.-H. Chen, "Overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task," *Proceedings of the NTCIR-6 Workshop Meeting*, Tokyo, Japan, 2007.

[7] B. Katz and B. Levin, "Exploiting Lexical Regularities in Designing Natural Language Systems," *Proceedings of the 12th International Conference on Computational Linguistics (COLING '88)*, Budapest, Hungary, 1988.

[8] K. Knight and J. Graehl, "Machine Transliteration," *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Somerset, New Jersey, 1997.

[9] M. Arbabi, S. Fischthal, V. Cheng, and E. Bart, "Algorithms for Arabic Name Transliteration", *IBM Journal of Research and Development, 38:2*, 1994.

[10] S. Wan and C. Verspoor, "Automatic English-Chinese Name Transliteration for Development of Multilingual Resources," *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, Canada, 1998.

[11] Y. al-Onaizan and K. Knight, "Machine Transliteration of Names in Arabic Text," *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, Pennsylvania, 2002.

[12] N. AbdulJaleel and L. Larkey, "Statistical Transliteration for English-Arabic Cross Language Information Retrieval," *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM '03)*, New Orleans, Louisiana, 2003.

[13] C.-J. Lee and J. Chang, "Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts Using a Statistical Machine Transliteration Model," *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts*, Edmonton, Canada, 2003.

[14] B. Xiang, J. Xu, R. Bock, I. Bulyko, J. Maguire, S. Matsoukas, A.-V. Rosti, R. Schwartz, R. Weischedel, and J. Makhoul, "The BBN Machine Translation System for the NIST 2006 MT Evaluation", *Proceedings of the NIST 2006 MT Workshop*, September 2006.

[15] B. Katz, G. Borchardt, and S. Felshin. "Syntactic and Semantic Decomposition Strategies for Question Answering from Multiple Resources", *Proceedings of the AAAI 2005 Workshop on Inference for Textual Question Answering*, Pittsburgh, Pennsylvania, 2005.

[16] B. Temulkuran, *Hap-Shu: A Language for Locating Information in HTML Documents*, Master's Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2003.

[17] G. Zaccak, *Wrapster: Semi-Automatic Wrapper Generation for Semi-Structured Websites*, Master's Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2007.