An Analysis of Question Processing of English and Chinese for the NTCIR 5 Cross-Language Question Answering Task

John Judge[1,2] Yuqing Guo[1,2] Gareth J. F. Jones[1,2,3] Bin Wang[4]
[1]School of Computing
[2]National Centre for Language Technology [3]Centre for Digital Video Processing
Dublin City University, Glasnevin, Dublin 9, Ireland
{john.judge,yuqing.guo,gareth.jones}@computing.dcu.ie
[4]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R.China
wangbin@ict.ac.cn

Abstract

An important element in question answering systems is the analysis and interpretation of questions. Using the NTCIR 5 Cross-Language Question Answering (CLQA) question test set we demonstrate that the accuracy of deep question analysis is dependent on the quantity and suitability of the available linguistic resources. We further demonstrate that applying question analysis tools developed on monolingual training materials to questions translated Chinese-English and English-Chinese using machine translation produces much reduced effectiveness in interpretation of the question. This latter result indicates that question analysis for CLQA should primarily be conducted in the question language prior to translation.
Keywords: *question analysis, question translation, cross-language question answering.*

1 Introduction

Question Answering (QA) systems are currently a major research field in applied natural language processing and related areas. The objective of a QA system is broadly to take a user's question of an information need expressed in natural language and seek an answer from a collection of documents. Depending on the specification of the system the questions may be from an unbounded range of subjects or from a specialized domain. The collection of documents from within which the answer is to be identified may range from everything available (e.g. from a web crawl) to a carefully selected set of trustworthy material (e.g. medical publications). Similarly the documents may vary from highly structured data from a single source to very heterogeneous and from multiple sources. The nature of the questions posed can vary from those with a single factoid answer to those requiring deep linguistic processing of documents to perform some comparative analysis of the information extracted from the documents to answer much more complex user information needs.

Regardless of the question type or information sources, a key issue for all QA systems is appropriate interpretation of the question. If a QA system is to answer questions accurately, it must accurately interpret the correct "meaning" of the question. The sophistication of the processing tools required depends on the linguistic complexity of the questions and the depth of analysis required to interpret the type of question entered. When developing QA systems for different languages the QA technologies, including the question analysis component, must be adapted to each language. The NTCIR 5 Cross-Language Question Answering (CLQA) task offers the opportunity to explore question processing on questions expressed in very different languages. Although the questions in this task are limited to those with named entity answers, we explore deep linguistic processing of questions in monolingual and cross-language environments. The motivation for this study is to develop language processing technologies that can ultimately be applied to more difficult QA and CLQA tasks. In this study we investigate the parsing and functional annotation accuracy of questions taken from English and Chinese data sets. Experimental results reported in this paper illustrate that differing levels of maturity in the linguistic resources used for question processing can impact on the accuracy of question interpretation.

An important element of CLQA is the exploration of the ablity of systems to accept questions in one language and find answers from a document collection in another language. A key question for such systems is how the language barrier between the question and document language should be crossed to provide the

"best" CLQA system. Best here most obviously refers to the accuracy of the answers, but might also include some consideration of the degree of coverage of the questions that can be accepted by the system. An interesting question for CLQA is to what extent the question should be processed in its original language and the intention of the question translated into a representation suitable for answering the question from the available documents, or the question itself translated prior to analysis and the QA system proceeding as a monolingual process in the document language. There is limited existing work in CLQA which begins to address this question [1][2]. This existing work attempts to identify question type, useful phrases and named entities in the question language, but does not explore the deep linguistic analysis required for detailed interpretation of the question. In this paper we begin to explore this issue by applying our question analysis methods to the output of standard machine translation resources for English-Chinese and Chinese-English. These experiments illustrate that the output of current machine systems is structurally very different from native text written in the output language and the results of question analysis using tools developed on natural language text is much reduced compared to corresponding monolingual questions.

This paper is organised as follows: Section 2 describes our question analysis methods for parsing and annotation of questions; Section 3 outlines the specific details of our question processing system; Section 4 details the question data sets used in our investigation; Section 5 gives the results of our monolingual and cross-language question analysis; and finally Section 6 describes the conclusions of our study and outlines directions for our further work on CLQA.

2 Question Analysis

Our question analysis system performs deep linguistic processing of the question in two stages: parsing and annotation.

2.1 Parsing

Parsing is the process of analysing (natural or formal) language into its component parts and describing how these relate to each other syntactically. For natural language input, parsing usually produces an output showing the lexical category of each of the words and the internal structure of the input (a parse tree and/or a more abstract dependency or logical form structure). Parsing natural language gives us useful information about the internal structure of sentences and phrases. Figure 1 shows a parse tree structure for a simple sentence.

Parsing sentences allows us to extract "deep" linguistic information about the sentence structure and

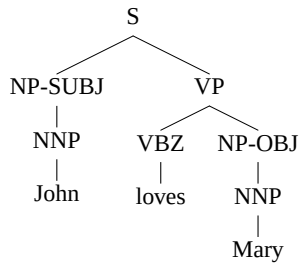

Figure 1. Example Tree

internal relationships like subject and object. These are key factors in disambiguating sentences which have similar surface forms, and as such can be crucial in disambiguating questions which have similar surface form, but are seeking different information. For example the questions "Who killed Harvey Oswald?" and "Who did Harvey Oswald kill?" have a similar surface form, but are seeking entirely different information as a response to the question. The first is seeking a named entity that can fulfill the subject role of the sentence, the second is seeking the object. Using parsing to identify these differences is an effective way of extracting this kind of information. Performing shallow processing of questions to identify expected answer types and named entities can often fail to capture subtleties of question interpretation of this type, and may lead to the return of quite incorrect answers or difficulties in selecting the correct answer from among multiple possibilities found in the available document set.

2.2 Annotation

Lexical Functional Grammar (LFG) [7] is a meaning based formalism which analyses sentences at a deeper level than syntactic parsing. LFG analysis identifies the main predicate of each clause (and subclause) and shows how functional roles such as subject, object, modifier and quantifier are fulfilled by various lexical items. LFG uses two levels of representation c(onstituent)-structure, which corresponds to the output of parsing, and f(unctional)-structure which represents functional roles and relations. This type of analysis is useful in that it is a more abstract representation of linguistic information than a parse tree structure. In addition, long distance dependencies, which are very common in interrogative sentences and fact seeking questions, are resolved in order to have a complete and correct f-structure analysis. This makes LFG analysis useful for QA tasks because it identifies the focus of the question and also which functional role (e.g. subject and object) the focus can fulfill. Figure 2 shows an f-structure for a simple sentence, and Figure 3 an f-structure for a question from the NTCIR 5 CLQA test set, the resolved dependencies are indicated by co-indexation. The f-structure shows that the

"John loves Mary"

$$\begin{bmatrix} \text{PRED} & \text{'love}\langle \text{SUBJ , OBJ} \rangle\text{'} \\ \text{TENSE} & \text{PRES} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'John'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Mary'} \end{bmatrix} \end{bmatrix}$$

Figure 2. Example F-Structure

"Who made Big Wonder?"

$$\begin{bmatrix} \text{PRED} & \text{'make}\langle \text{SUBJ , OBJ} \rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \boxed{1} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Big Wonder'} \end{bmatrix} \\ \text{FOCUS} & \begin{bmatrix} \text{PRED} & \text{PRO} \\ \text{PRONFORM} & \text{'who'} \end{bmatrix} \boxed{1} \end{bmatrix}$$

Figure 3. Example F-Structure for a Question

focus of the question is the subject of the main predicate of the sentence "make".

LFG analysis thus provides valuable information for the detailed interpretation of complex questions which can potentially form a significant component in answering them correctly.

3 Question Processing System

For this investigation of question processing we use Bikel's parser [3] to generate c-structure trees, and the automatic annotation algorithm of [4, 5] to generate f-structures from the c-structure trees using a simple pipeline architecture, as shown in Figure 4.

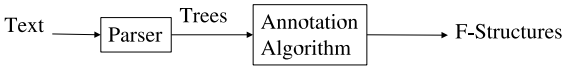

Figure 4. Pipeline Architecture

The parser in each language is trained on a standard training set for the language, the Wall Street Journal sections 2-21 of the Penn-II Treebank [8] for English, and the Chinese Treebank-II for Chinese [9]. The annotation algorithm of [4, 5] is unchanged from this previous work.

There are a number of differences between the linguistic resources used for English and Chinese. Firstly, the training corpora for the parser are unevenly matched in terms of size. The Wall Street Journal sections of the Penn-II Treebank which are used to train the English parser contain approximately 40,000 sentences. The Chinese Treebank-II on the other hand contains only 4,000 sentences[1]. The size of the training corpus can adversely effect parser performance if the corpus is of insufficient size to contain enough data to allow the grammar to generalise over unseen data. Secondly, the LFG annotation algorithm used to generate the f-structures, is developed monolingually, and at present the Chinese modules for the annotation algorithm are relatively new and hence have not been developed to the same extent as the English equivalent. Furthering the quality of these tools is part of ongoing work in our research group.

4 NTCIR 5 CLQA Question Sets

In Section 5 we describe a set of experiments designed to investigate the effectiveness of the analyses described above for questions in the NTCIR 5 CLQA Question Sets, and also to determine the effect of automatically translating the questions. In order to be able to perform evaluations of these kind, we need a "gold standard" against which to evaluate them. The gold standard set contains analyses of the test set which are deemed to be correct as verified by a human annotator.

Due to the time consuming and difficult nature of creating such a gold standard for the entire NTCIR 5 CLQA question test sets, we randomly selected a subsection of 50 translationally equivalent questions in both English and Chinese taken from the NTCIR 5 CLQA test sets and manually verified these as gold standards for our evaluations. We believe that performance on this subset of the test set to be representative of performance on the full set.

4.1 C-Structure Gold Standard

We created a gold standard of c-structure trees from the 50 raw questions by first passing them to a state-of-the-art parser [3] and then hand correcting the mistakes made by the parser. Figure 5 shows an example c-structure tree as output by the parser (a) and then after hand correction for the gold standard (b).

4.2 F-Structure Gold Standard

Previous work on automatic f-structure annotation in [4, 5] has shown that given a c-structure analysis of

[1] A larger Chinese Treebank (CTB4) exists, however work on Chinese question analysis is at an early stage of development and we have yet to scale up to this larger data set.

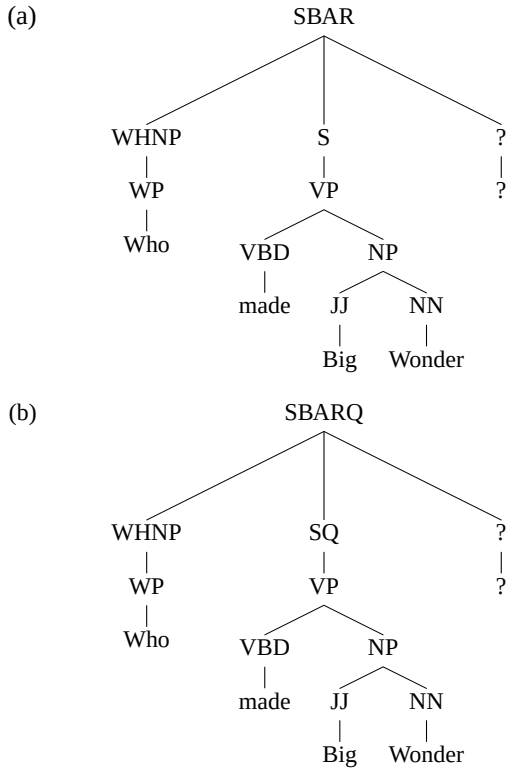

Figure 5. Example trees before and after hand correcting

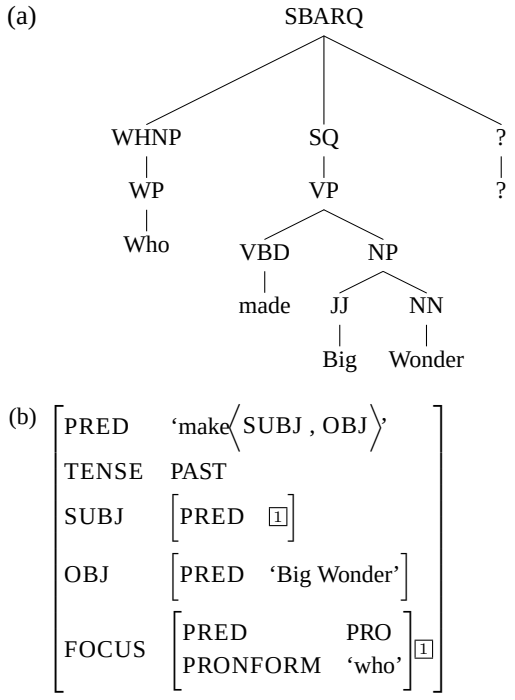

Figure 6. A gold standard tree and corresponding gold standard f-structure.

high enough quality the automatic LFG annotation resources of [4, 5] will produce f-structures of equally high quality. In light of this we generate the gold standard f-structures automatically from the gold standard c-structure trees which have been hand corrected and are deemed to be correct. This process is used to generate gold standards for English and Chinese f-structures. Figure 6 shows the example gold standard tree and its corresponding gold standard f-structure.

5 Experimental Investigation

In this section we describe results for our analyses of English and Chinese questions. We first explain the evaluation metrics used and then give monolingual and cross-language question analysis results.

5.1 Evaluation Metrics

In evaluating the quality of analysis of c-structure parsing and f-structure analysis we use precision and recall calculated on tree constituents and functional dependencies. In the c-structure analysis the constituents evaluated are the nodes of the parse tree, and in f-structure analysis the dependencies are evaluated. In both types of evaluation the constituents/dependencies are compared to the gold standard version of the tree/dependencies which has been hand checked to ensure correctness. Given an analysis P and a corresponding gold standard analysis T, precision, recall and f-score (the harmonic mean of precision and recall) are calculated as shown below.

$$precision = \frac{number\ of\ correct\ constituents\ in\ P}{number\ of\ constituents\ in\ P}$$

$$recall = \frac{number\ of\ correct\ constituents\ in\ P}{number\ of\ constituents\ in\ T}$$

$$f\ score = \frac{2 \times precision \times recall}{precision + recall}$$

5.2 Monolingual Question Analysis

5.2.1 English Questions

	Precision	Recall	F-Score
Trees	80.19	80.76	80.97
F-Structures	77.83	88.73	82.92

Table 1. English Question Analysis Results.

Table 1 gives results of the evaluation of our 50 English question test set. The analysis of the questions is quite good with an f-score in excess of 80% for c-structure parsing. Correspondingly the scores for the

f-structure analysis are high, with an f-score also over 80%. These results are comparable to results from earlier experiments for a similar analysis of questions taken from the ATIS corpus [6].

5.2.2 Chinese Questions

	Precision	Recall	F-Score
Trees	57.03	60.86	58.88
F-Structures	74.97	76.99	75.97

Table 2. Chinese Question Analysis Results.

Table 2 shows results for the analysis of Chinese c- and f-structure analysis for the 50 question test set. The f-score for the Chinese c-structures is considerably lower than for the corresponding English analysis, the f-structure analysis is also lower than the corresponding English result. This can be attributed to two factors.

First, the relative difference in size between the English and Chinese training corpora is likely to have affected the quality of the grammar and hence the evaluation results. The development of a set of high quality grammars for multiple languages, including Chinese and Japanese, is a current focus of a research project in our group. The impact of these improved resources on question analysis will be explored as these resources become available.

Second, the Chinese Treebank-II is modern simplified Chinese Mandarin, unlike the traditional Chinese Mandarin of the CLQA data set. This introduces issues of differences in both the character sets and grammars of the parser training set and question test set.

Some examples of problems that we observed from manual analysis of the questions are as follows:

- Chinese Treebank-II is a small corpus and taken from newswire, most sentences are declarative, whereas the test set are questions. Some of frequent interrogatives in the test set are not included in Chinese Treeback-II, such as the interrogative determiner "哪(which)".

- There are also notable language differences between modern simplified Chinese and traditional Chinese. For example, traditional Chinese has many mono-syllable words which derive from ancient Chinese. Those words have almost been abandoned or have evolved into other di-syllable or multi-syllable words in contemporary simplified Chinese.

For example, English:

What was the destination of the inaugural flight departing from Haneda Airport 's new runway B?

Traditional:

羽田 機場 新建 B 跑道 的 首航 班機 之 終點站 為 何 ？

Simplified:

羽田 机场 新建 B 跑道 的 首航 班机 的 终点站 是 哪里 ？

- Though the primary differences between the forms of Chinese are at word level, we identified a few at sentence level. For example, the collocation of "分隔" and "之间" would not be seen in contemporary simplified Chinese grammar, but is still used in traditional Chinese.

For example, English:

What is the name of the river that separates North Korea from China?

Traditional:

分隔(separate)中國與北韓之間(between)的界河叫什麼名字？

Simplified:

中国与北韩之间的界河叫什么名字？

and

分隔中国与北韩的界河叫什么名字？

The obvious solution to these issues would be build a question analysis module for the NTCIR 5 CLQA task using a tree bank constructed for traditional Chinese. Such resources are availble for research in this area, and we are in the process of acquiring a suitable training resource[2] to scale up our system.

5.3 Analysis of Translated Questions

One of the objectives of our work is the development of effective CLQA technologies. As part of our investigation in this section we report results for the analysis of the output of translating our test question set using standard machine translation (MT) systems. The questions were translated from English to Chinese and Chinese to English, using the Systran[3] online MT system. The MT output was analysed in the same way as the untranslated monolingual questions, and evaluated against the gold standards as before.

5.3.1 English-Chinese Translation

Table 3 shows an analysis of the MT system's Chinese translation of the 50 English questions. It is evident that there is a very significant reduction in parse accuracy relative to the Chinese monolingual result

[2] http://turing.iis.sinica.edu.tw/treesearch
[3] http://www.systransoft.com/

	Precision	Recall	F-Score
Trees	9.22	10.54	9.67
F-Structures	16.28	18.12	17.5

Table 3. Analysis of English to Chinese MT output.

shown in Table 2. The deeper f-structure analysis has also suffered with an f-score of 17.5, compared to the Chinise monolingual score of 75.97.

5.3.2 Chinese-English Translation

	Precision	Recall	F-Score
Trees	6.60	6.46	6.37
F-Structures	20.93	24.36	22.51

Table 4. Analysis of Chinese to English MT output.

Table 4 shows the breakdown for the analyses of the output of the MT system's translations of the 50 Chinese questions into English. The c-structure analysis has suffered considerably with an f-score of only 6.37 compared to 80.97 in the English monolingual analysis. The f-structure evaluation has also suffered a loss of accuracy, though to a lesser extent, with an f-score of 22.51.

The Chinese-English evaluation c-structure f-score is notably lower than the English-Chinese evaluation. This, we believe, can largely be attributed to the poor word order in the output of the MT system. This is a greater problem for English where the word order is not as free as in Chinese. Figure 7 shows an example taken from the MT output, and its corresponding English gold standard (GS) equivalent.

(MT) The South Korea biggest automobile manufacturer is?

(GS) What is the name of the South Korea's largest automaker?

Figure 7. Example Chinese to English MT vs GS English question.

The finding that both the English and Chinese translated output have very poor performance relative to the monolingual question suggests that this is not due to the available analysis resources, but rather the form of the translated output. Analysis of the translated questions indicates that although the translated output often has the correct or appropriate words and local phrases, there are often problems with exact word order for English and Chinese which the grammar of the question analysis parser is not able to process suitably. We have noted in previous work on the processing of the English ATIS test set questions that high quality parse trees are crucial to achieving high quality f-structures [6].

Evaluation of MT systems tends to focus on the accuracy of local translation of word groups rather than the type of long distance dependencies that are crucial to the correct understanding of questions, and it is thus possible that current MT evaluation metrics do not give a useful indication of the usefulness of individual MT systems for applications requiring deep understanding of the material to be translated, such as question interpretation.

This suggests that for CLQA as much question analysis as possible should be done in the source language, and the interpreted output translated in some way to the document language. For similar languages the f-structure will have a similar form, and this translation may be fairly simple. However, for very different languages, as is the case for English and Chinese, this translation process will be much more complicated.

A possible alternative approach to addressing this issue would be to train a grammar on an annotated corpus of output from the MT system to be used for the translation, but for this to be effective we would have to be sure that the translation output would be produced consistently according to some reliable rule set. Of further concern is the issue that the grammar would have to be adapted for any change to linguistic functionality of the MT system or possibly replaced entirely if a new MT system were to be introduced. There is of course the issue of the cost of manually annotating such a training corpus.

6 Conclusion and Further Work

The results of our analysis of English and Chinese questions from the NTCIR 5 CLQA test set show that we are able to build high quality f-structures for the monolingual English questions, but that the small size of our Chinese training set and linguistic mismatch between the training and test question sets mean that the performance of Chinese questions is currently lower. We are currently working on improving our Chinese grammar and annotation system, and results of question analysis should improve as these become available.

We have shown in previous work that the accuracy of question processing can be increased for questions from the English ATIS test set by including a significant number of annotated representative questions in the training set used to train the c-structure parser [6].

It would be interesting to investigate this for both English and Chinese for the NTCIR 5 CLQA test set. In particular it would be interesting to see the extent to which including an annotated question set might mitigate the problems of linguistic mismatch between our Chinese training set and the CLQA test questions.

Analysis of results for questions translated using standard MT are very poor indicating that the output of current MT resources is not suitable for the analysis of questions in CLQA systems. It is possible that such systems may be suitable for simple named entity questions not requiring detailed interpretation, but this will not be an effective approach as question complexity increases. Rather our results indicate that we should perform analysis of the question prior to translation and that the question should be matched to potential answers at the level of the f-structures. This process is far from straightforward, particularly for languages which differ significantly such as English and Chinese, and this topic will form an ongoing area of our further work.

Results from the numerous Cross-Language Information Retrieval (CLIR) evaluations show that MT, and indeed shallow statistical and dictionary-based methods, can cross the language barrier effectively for document retrieval. Thus, it should not be assumed that an effective CLQA system should translate the question in the same way both with respect to retrieving documents which may contain the correct answer and identifying the answer in documents in the target language. Investigation of this aspect of the design of effective CLQA systems will also form a component of our further work.

References

[1] S. Sekine and R. Grishman. Hindi-English Cross-Lingual Question-Answering System. *ACM Transactions on Asian Language Information Processing*, 2(3):181-192, September 2003.

[2] C. Peters et al. (editors) *Proceedings of the CLEF 2005 Workshop*, Vienna, September 2005.

[3] D. M. Bikel. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of HLT 2002*, pages 24–27, San Diego, CA, 2002.

[4] M. Burke, A. Cahill, R. O'Donovan, J. van Genabith, and A. Way. The Evaluation of an Automatic Annotation Algorithm against the PARC 700 Dependency Bank . In *Proceedings of the Ninth International Conference on LFG*, pages 101–121, Christchurch, New Zealand, 2004.

[5] A. Cahill, M. Burke, R. O'Donovan, J. van Genabith, and A. Way. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 320–327, Barcelona, Spain, 2004.

[6] J. Judge, A. Cahill, M. Burke, R. O'Donovan, J. van Genabith, and A. Way. Strong Domain Variation and Treebank-Induced LFG Resources. In *Proceedings of the Tenth International Conference on LFG (LFG05)*, Bergen, Norway, July 2005.

[7] R. Kaplan and J. Bresnan. Lexical Functional Grammar, a Formal System for Grammatical Representation. In J. Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press, Cambridge, MA, 1982.

[8] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[9] N. Xue, F.-D. Chiou, and M. Palmer. Building a Large-Scale Annotated Chinese Corpus. In *Proceedings of the 19th. International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, August 2002.