An Investigation of Question Translation for English–Chinese Cross-Language Question Answering

Ying Zhang[1], Gareth J. F. Jones[1,2], Sen Zhang[3], Bin Wang[3], Yuqing Guo[2], Yanjun Ma[2]

[1]Centre for Digital Video Processing
[2]National Centre for Language Technology
Dublin City University, Glasnevin, Dublin 9, Ireland
{ying.zhang,gareth.jones,yuqing.guo,yanjun.ma}@computing.dcu.ie

[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R.China
{zhang,wangbin}@ict.ac.cn

Abstract

Question-answering (QA) is a next-generation search technology which aims to provide answers to a user's question from a collection of documents. Cross-Language QA (CLQA) extends this paradigm to answering questions from a collection in a different language to the question itself. The accuracy with which a CLQA system answers questions depends on the QA system and translation between the question and the information source. We report results from an evaluation of English–Chinese CLQA comparing question translation using standard machine translation systems and extended translation incorporating web mining to enhance the translation dictionary against a baseline of monolingual Chinese QA. Results from these experiments show that our noun phrase recognition and translation techniques lead to a significant improvement in CLQA effectiveness. Moreover, the syntactic form of a question can be impaired during query translation, and thus potentially degrades the overall CLQA system performance.

Keywords: Cross-language question answering, Named entity extraction, Web mining.

1 Introduction

In a few short years search engines such as *Google* have become a ubiquitous tool in many people's lives. However, current search technologies represent only the start of what may soon be possible for exploiting the huge amounts of digital information that are increasingly becoming available on the internet and elsewhere. One next-generation search technology which is currently the focus of considerable research is Question Answering (QA), which aims not just to locate a document containing the answer to a question, but to actually extract the answer itself and return this to the user directly. Current search engines face the challenge of locating documents which are relevant to the searcher's query, and do this by matching words in the query against those in the documents using a range of algorithms to rank documents most reliably. QA systems face the additional challenges of "understanding" the question, at least to some degree, to determine what the user wants to know, and analyzing documents to locate potential answers to this question.

An ongoing research area in search technology is cross-language information retrieval (CLIR) where a query in one language is used to search for relevant documents in a different language, e.g. using an English query to search a Chinese document collection. The primary challenge of CLIR beyond standard single language monolingual search is crossing the language barrier to reliably match the words in the query and the document. An obvious approach is to apply a standard machine translation (MT) system to translate the query. However, search key terms are often rare or domain-specific words outside the vocabulary of an MT system. Short queries (often two or three words) may make application of an MT

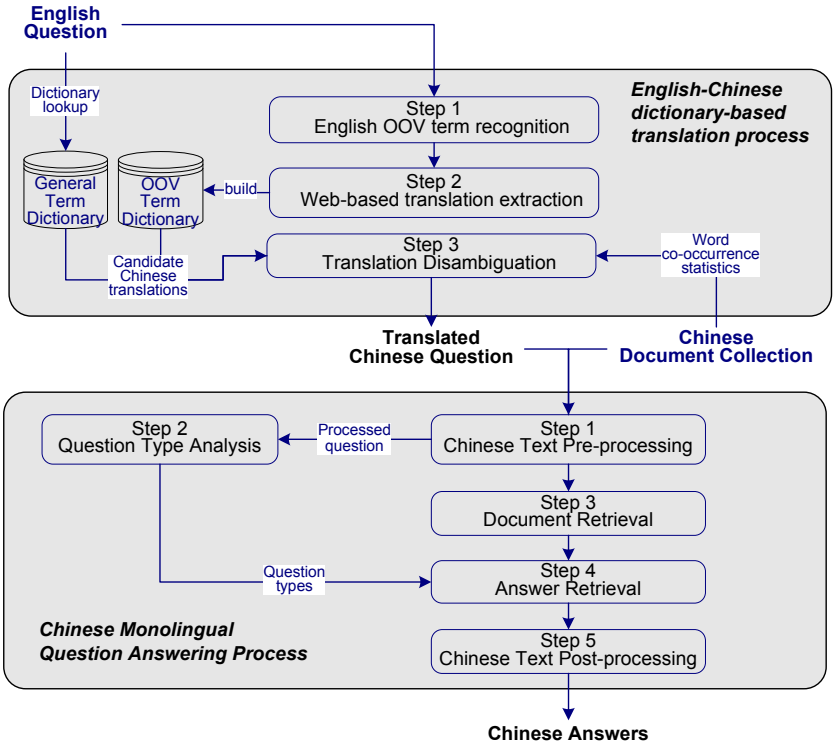

Figure 1: Flow chart of English–Chinese question answering process.

system, which is designed for natural language texts, unreliable due to lack of context and linguistic structure. Many techniques to address these problems have been explored in recent years, and CLIR research is still ongoing. Interest is now developing into the still more challenging topic of cross-language question answering (CLQA). This faces the same problems as monolingual QA and CLIR, but the additional challenge of reliably translating sufficient detail of the question to enable it to be interpreted correctly. In previous work we demonstrated that applying a standard MT system to question translation between Chinese and English does not produce an accurate rendering of the structure of the question [Judge et al., 2005]. In order to explore English–Chinese CLQA we first developed a monolingual Chinese QA system for the NTCIR-6 workshop [Zhang et al., 2007]. In this paper we extend our previous work to explore question translation for English–Chinese CLQA. In particular we investigate the development of reliable translation of words and phrases unknown to an MT system by using web mining of online resources such as Wikipedia.

This paper is structured as follows: Section 2 outlines the architecture of our CLQA system, Section 3 reviews our monolingual Chinese QA system, Section 4 describes our translation strategies, Section 5 gives details of the experimental investigation and results, and finally Section 6 concludes the paper.

2 System Architecture

Our English–Chinese CLQA system uses a two-phase process as shown in Figure 1: first, English questions are translated into Chinese; second, for each translated Chinese question, the answers are extracted from the Chinese document collection using our Chinese monolingual QA system. The Chinese QA system adopts a document retrieval followed by answer extraction approach. First, we retrieve short document passages that are potentially relevant to a question and may contain the answer to the question; and then, content analysis methods are used to mark and obtain the most likely answer from the retrieved passages. The stages of this system are described in more detail in the following sections.

3 Monolingual Chinese Question Answering

3.1 Pre-processing and Post-processing

In the NTCIR-6 CLQA task explored in this paper both the Chinese questions and documents are encoded with BIG5 ordinarily used in Taiwan [Sasaki et al., 2007]. Since our language processing is based on the GBK encoding used in the mainland of China, we first convert all text from BIG5 into GBK encoding using *Textpro*. The data is then split into short document passages for retrieval. These short documents form the basic retrieval unit in the QA system, once they are retrieved in response to a question, the QA system attempts to locate the answer within these passages. Documents are divided into passage documents using the HTML markup tags in the documents. The ICTCLAS tool is used to automatically segment the data into Chinese words, assign a Part-of-Speech (POS) tag to each word and recognise named entities. Although ICTCLAS is trained on simplified Chinese news articles from the *People's Daily*, we found that ICTCLAS can generate reasonable segmentation results and named entity labels for the transformed document collection and question set. In the post-processing stage, answers are converted back to BIG5-encoding to form the final output.

3.2 Question Type Analysis and Classification

In order to answer questions accurately, the returned answer should correspond to the type of answer expected from the question. For the CLEF-6 CLQA task all answers are of type Named Entity, but still the answer must be of the appropriate entity type. We therefore use pattern matching with heuristic rules to classify the answer type of each question as one of a set of predefined types — PERSON, LOCATION, ORGANIZATION, DATE, TIME, NUMEX, MONEY, PERCENT, and ARTIFACT.

3.3 Document Retrieval

The Lemur toolkit is used to perform passage retrieval. Lemur supports the use of a number of different information retrieval ranking models, after informal experimentation we decided to use the Okapi BM25 model for document retrieval [Robertson et al., 1995]. Some stop items such as interrogatives and other common stop words were eliminated from questions. We have also explored giving different weights to certain Chinese terms appearing to be more important, such as proper nouns and entity names; however, the results were not clearly improved.

3.4 Answer Extraction

We extract candidate answers for each question from the top-20 ranked retrieved passages using various strategies based on the question type. Heuristic rules together with ICTCLAS are used to extract the candidate answers from the passages. ICTCLAS is able to recognize proper nouns (including personal names, location names, and organization names), temporal words (including date and time), numeral words (money, numex, percent) and ordinary artificial nouns; however it is not sufficient for discrimination between different numerical types. For example, ICTLAS recognizes NUMEX, PERCENT, and MONEY as numeral words, but cannot distinguish them. Future work will focus on extending the capabilities of ICTLAS. Finally, the number of the occurrences of each candidate answer is counted. The candidate answer with the highest occurrence frequency is selected as the most likely answer to the question.

4 English–Chinese Query Translation

In this section we describe the components of our English-Chinese translation process. There is an ongoing debate in CLIR research regarding whether to translate search queries into the language of the documents, or documents into the language of the query. For this study we focus on translation of a question into the document language, and limit our discussion to query and question translation issues.

4.1 Enhanced MT-based Query Translation

Machine Translation (MT) based query or question translation uses an existing MT system to provide automatic translation. Using MT systems for query translation is quite popular in CLIR when such a system is available for the particular language pair. In the NTCIR-5 workshop English–Chinese and Chinese–English CLIR tasks [Kishida et al., 2005], most participants used the Systran MT system in some form for parts of their experiments. While MT systems can provide reasonable translations for general language expressions, they may not be sufficient for topical entities such as personal names, organization names, movie names, place names, etc.

To enhance the standard MT-based query translation, we implemented automatic English out-of-vocabulary (OOV) term recognition and web-based translation procedures (explained later) to translate the English terms previously undetected or untranslated before the MT software operations.

4.2 Dictionary-based Query Translation

Due to the increasing availability of machine readable dictionaries and the linguistic limitations of MT systems, much research effort in CLIR has been focused on dictionary-based methods for query translation. Using this strategy, queries are translated by looking up words in a bilingual dictionary and using some or all of them as the translated query.

We compiled an English–BIG5 general-term translation dictionary using four dictionaries for the English–Chinese query translation task: an English–Chinese wordlist and a Chinese–English wordlist from LDC, a Chinese–English wordlist from CEDICT, and an English–Chinese wordlist from Britannica Online. [Kwok, 2000] showed that the Chinese–English wordlist can be considered as both a phrase and word dictionary for English–Chinese cross-language retrieval, and is preferable to the English–Chinese version in terms of phrase translations and word translation selection. Our general-term translation dictionary contains $179,271$ entries including $59,371$ multi-word phrases that were used for English phrase identification and translation.

As shown in Figure 1, English OOV term (single-word and multi-word terms) recognition is the first step of our query translation process. We then apply our web-based OOV translation techniques to extract Chinese translations and add new terms into the translation dictionary. Once candidate query term translations are collected, we use a disambiguation technique to determine the most appropriate Chinese translation for each English query term. Our dictionary-based query translation results are shown as E–C–DICT in Table 1.

OOV term recognition

Single-word English OOV terms can be recognized easily, since they are either present in the translation dictionary or not. However, if a multi-word expression is missing from the phrase dictionary, it will ordinarily be translated word by word. This is often inappropriate since the meaning will not be accurately conveyed in the translation. Phrase recognition refers to the identification of a group of words with a special meaning when they co-occur. For example "great leap forward","science and technology", "Columbine high school", etc. It is much more appropriate to identify such multi-word expressions as an OOV term and translate them as a phrase.

Our multi-word OOV term recognition process builds on a base noun phrase (baseNP) chunking module in which a sentence is divided into a sequence of non-overlapping, non-recursive segments of text, representing specific grammatical categories. Our chunking process of phrase recognition involves two phases: bracketing and labeling. We first identify the chunk boundaries in a sentence, and then classify the chunks into appropriate grammatical classes. In our implementation, these two phases are combined together in the chunking module. [Ramshaw and Marcus, 1995] first introduced the machine learning method to the baseNP chunking task. They tagged each word with a "chunk tag" in the IOB format, where words are inside a chunk (I), outside a chunk (O), and at the beginning of a consecutive chunk (B). Various machine learning techniques have been developed for noun phrase chunking [Sang and Buchholz, 2000], such as Support Vector Machines (SVMs), Hidden Markov Models, Memory Based Learning and Conditional Random Fields. Among all these techniques, SVMs provided

the highest accuracy for English baseNP identification [Kudo and Matsumoto, 2001], and thus are employed in our experiments. We used chunks converted from the Wall Street Journal portion of the Penn Treebank [Marcus et al., 1994] as our training data. All the experiments are carried out using the YamCha system for SVM-based chunking process. 90% of the Penn Treebank were used as a training set and the remaining 10% as a test set to evaluate the performance of our chunk processing. The evaluation showed that this approach has a precision of 96.5%, a recall of 96.5%, and a F-score of 96.5%.

In summary, we tokenised the 150 English question sentences from the NTCIR-6 CLQA task and tagged them with a maximum entropy POS tagger [Ratnaparkhi, 1996]. The tagged sentences were then passed into the YamCha system. As a result 597 English phrases in total were identified from the all question sentences. There were 294 English phrases that could not be found in our general-term translation dictionary and passed into our web-based translation extraction process.

Translation extraction using the Web

In this section, we describe our process for automatically extract Chinese translations of English OOV terms from the web. In formulating our approach, we combined two methods — wikipedia-based extraction and search-based extraction — to improve the translation accuracy.

Wikipedia-based extraction As a multilingual hypertext medium, Wikipedia has been proved to be a valuable new source of translation information [Adafre and de Rijke, 2006]. Wikipedia is structured as an interconnected network of articles, in particular wikipedia pages titles in one language are linked to a multilingual database of corresponding terms. Unlike the web, most hyperlinks in wikipedia have a more consistent pattern and meaningful interpretation. For example, the English wikipedia page http://en.wikipedia.org/wiki/Formosa_Plastics contains a hyperlink to its counterpart written in Chinese http://zh.wikipedia.org/wiki/台塑企業, where the basenames of these two URLs ("Formosa Plastics" and "台塑企業") is an English–Chinese translation pair.

To utilize this multilingual linkage feature, we implement a three-stage process to extract English–Chinese translations from the wikipedia automatically.

1. Given an identified noun phrase consisting of n words ($e_1 e_2$ ... e_n), we use the whole phrase as a query to search the English wikipedia and save the HTML source of the returned document as a local file using the following command:

   ```
   lynx -source
   http://en.wikipedia.org/wiki/Special:Search?search=e_1+e_2+..+e_n&go=Go
   > local_output_file
   ```

2. We then extract the URL zh.wikipedia.org/wiki/Chinese_Term which appears in the hyperlink to the Chinese wikipedia page using the following pattern

   ```
   <a href="http://zh.wikipedia.org/wiki/.*">中文</a>
   ```

3. If such a URL exists, we select its basename as the Chinese translation of the English term ($e_1 e_2$ ... e_n), and add this translation pair into our English–Chinese dictionary.

Using this method, we were able to find 101 English–Chinese translation pairs. These translation pairs were then added into our OOV-term translation dictionary.

Search-based extraction When new terms, foreign terms, or proper nouns are used in Chinese web text, they are sometimes accompanied by the English translation in the vicinity of the Chinese text. By collecting a sufficient number of such instances for a given English term and applying statistical techniques, we are able to infer its Chinese translation(s) with reasonable confidence.

We use Google to fetch up to 300 Chinese BIG5 documents using the English noun phrases identified previously as the queries. For each returned document, only the title and the query-biased summary (the text snippet shown to the user in the ranked output) are extracted and then filtered by an HTML parser and segmented using the Chinese punctuation delimiters to collect lines containing English strings.

Chinese text within various types of Chinese quotation marks is observed sometimes to be followed by English strings within brackets. For example: 「口袋怪獸」（pocket monster）, 「神奇寶貝 (Pocket Monster), 《美日安保條約》（Japan-US Security Treaty）, 「大浦洞一號」 (Taepodong-1), 《世界末日》（Armageddon）, 「中華衛星一號 (ROCSAT-1)」, 《紅色角落》（Red Corner）, 《火並時速》(Rush Hour), 「尖峰時刻（Rush Hour）」, etc. Our previous experience in web-based translation extraction [Zhang et al., 2005] showed that that 98% of these were correct translation pairs. The high accuracy is both because this format is a convention commonly used to indicate translations and also because the use of quotation marks eliminates problems associated with Chinese word segmentation. If we have added more than one translation to the translation dictionary (for example, both "火並時速" and "尖峰時刻" are added as the Chinese translations of the English term "Rush Hour"), we use our disambiguation technique to select the most appropriate alternative in the given context.

For the instances which lack explicit delimiters (such as punctuation marks) in the Chinese text, we collect the frequency and the length of every possible Chinese string (up to 20 Chinese characters) occurring adjacent (before and after) to the English OOV term. The Chinese substring with the highest weight, which is computed based on frequency and length analysis, is selected as the translation of the English OOV term [Zhang and Vines, 2004]. In total, we extracted 164 translation pairs using this method. The extracted translation pairs were then added into our OOV-term translation dictionary for query term dictionary lookup.

Translation disambiguation

Translation ambiguity is a frequent cause of failure of dictionary-based translation due to the fact that many terms in one language can be translated into another language in multiple ways, and sometimes the alternate translations have very different meanings. For example, the English term *director* can be translated as "指揮者"(conductor), "導演"(movie director), "董事"(a board of directors), and "官員"(official) in Chinese, the choice of Chinese translation depends on the context in which *director* occurs.

We use statistics obtained from the NTCIR-6 Chinese corpus for English–Chinese translation disambiguation. Given an English query and a set of Chinese candidate translations, each candidate translation is a sequence of Chinese terms. Our idea is to estimate the likelihood of each sequence of Chinese terms using a probability model, and select the one with the maximum likelihood among all possible candidate translations as the most appropriate translation of the given English query. Our disambiguation technique is based on a Markov model; such models have been used widely for probabilistic modelling of sequence data.

$$P(t_1, t_2, t_3 ... t_n) = P(t_1) \prod_{a=2}^{n} P(t_a | t_{a-1})$$

To compute the probability of a sequence of terms, we need to calculate the values of $P(t)$, the probability of term t, and $P(t|t')$, the probability of t in the context of t', as follows:

$$P(t) = \frac{f(t)}{N}\ ,\ P(t|t') = \frac{P_w(t,t')}{\sum_{t''} P_w(t'',t')}$$

where $f(t)$ is the collection frequency of term t, N is the number of terms in the document collection, and $P_w(t,t')$ is the probability of term t' and term t co-occurring within a window of size w. The variation in window size used to collect word association information has a small effect on the outcome, with $w = 4$ producing the best results. The zero-frequency problem arises in the context of probabilistic language models, when the model encounters an event in a context in which it has not been seen before. Smoothing provides a way to estimate and assign the probability to that unseen event. We used the following absolute discounting and interpolation formula, which applies the smoothing method proposed by [Federico and Bertoldi, 2002]. In this method,

$$P(t|t') = \max\left\{\frac{f_w(t,t') - \beta}{N}, 0\right\} + \beta P(t) P(t')$$

where $f_w(t, t')$ is the frequency of term t' and term t co-occurring within a window size w. The absolute discounting term β is equal to the estimate $\beta = n_1/(n_1 + 2n_2)$, where n_k is the number of terms with collection frequency k.

5 Experiments

This section first describes the NTCIR-6 CLQA test collection used in our experiments, summarises our experimental setup, then gives results and analysis of our different methods for question translation.

5.1 Test Collection

NTCIR is a series of evaluation workshops organised in Japan on an 18 months cycle for the evaluation of Asian language information access technologies. The fifth and sixth NTCIR workshops have offered a CLQA task between various language pairs. In this paper we base our investigation on use of the Chinese–Chinese (C–C) and English–Chinese (E–C) CLQA tasks from NTCIR-6. The full details of this task are given in [Sasaki et al., 2007].

Answers to questions must be extracted from a set of documents taken from United Daily News, Economic Daily News, Min Sheng Daily, United Evening News, Star News from $1998 - 1999$. Each question has only one answer and is restricted to being a Named Entity: proper noun, such as the name of a person, an organisation, various artifacts, and numerical expressions, such as money, size, date, etc. 150 test questions were provided for the task, and after completion of the task the answers to the questions were made available by the organisers. Results shown are the number of answers which are correct out of 150 for each method. The task requires that the system both supplies the answer and the document from which it was taken. Answers are either "Right" (R), the answer is correct, and the document from which is taken contains the answer, or "Right and Unsupported" (R+U), the answer is correct, but cannot be inferred directly from the indicated document.

5.2 Experimental Setup

The following three runs were performed in our English–Chinese CLQA experiments:

- E–C–MT: To provide a baseline for our CLQA results, we translated each of the English questions using the Systran free online translation service.
- E–C–MT+OOV: To augment the standard MT-based query translation, we identified and translated the English OOV terms using the translations extracted from the web before the MT translation. To translate an English query, the first step is to identify the English OOV terms and replace them using all Chinese translations via an OOV-term dictionary look-up. We then used Systran to translate the re-formulated questions. The Chinese terms appearing in a re-formulated question were unaltered during the Systran translation process.
- E–C–DICT: English questions were translated using the disambiguation technique combined with the web-based translation extraction technique described in Section 4.2.

5.3 Results and Discussions

The evaluation results are shown in Table 1. Our OOV term recognition and translation techniques lead to a significant improvement in CLQA effectiveness. We note that we get higher correctness for those questions asking about names of person, location and organization, while no correct answers were found for numex, percent and money. This **is** mainly due to the fact that ICTCLAS only performs well in recognizing personal names, location names and organization names, as described in Section 3.4. The high correctness for date and time are perhaps due to the fact that there is no significant difference between these two concepts. Further it is possible that tools trained on Simplified Chinese corpora have some deficiencies when attempting to process the Traditional Chinese documents and queries.

QType	QID	E–C–MT		E–C–MT+OOV		E–C–DICT		C–C	
		R	R+U	R	R+U	R	R+U	R	R+U
ARTIFACT	7	0	0	0	0	0	0	1	1
DATE	39	0	1	1	2	5	7	12	13
LOCATION	16	1	1	0	0	0	1	8	9
MONEY	8	0	0	0	0	0	0	0	0
NUMEX	11	0	0	0	0	0	0	0	0
ORGANIZATION	16	0	0	0	1	0	0	5	5
PERCENT	4	0	0	0	0	0	0	0	0
PERSON	47	1	2	4	12	4	7	16	22
TIME	2	0	0	0	0	0	0	0	0
Number of correct	150	2	4	5	15	9	15	42	50
Accuracy (%)	—	1.33	2.67	3.33	10	6	10	28	33

Table 1: NTCIR-6 English–Chinese CLQA and Chinese monolingual QA evaluation results.

Interestingly, our techniques sometimes produced translations that might be considered more correct than the provided translation. For example, we translated the English question "When was Pokemon invented" as "何时 (what time) 神奇宝贝 (Pokemon) 发明 (invent)", which is arguably more correct than the original monolingual Chinese question "神奇宝贝 (Pokemon) 是 (is) 什么 (what) 时候 (time) 被创造 (invent)". The Chinese terms "发明" is synonymous with "创造", but more appropriate in the given context. These two translations were both accurately classified as DTAE-type questions. However, our POS tagging tool incorrectly marked the Chinese term "发明" as a noun, whereas the Chinese term "创造" was correctly marked as a verb. This is because the sentence structure was changed during the query translation process, and this alteration affected the POS interpretation of the translated Chinese text. Due to the fact that the verb is defined as a starting pattern for DATE-type answers in our heuristic rules, our answer retrieval component failed to retrieve any valid answer for the translated Chinese question.

6 Conclusion

In this paper, we have looked at the English OOV problem as it applies to English–Chinese CLQA. We adopted a baseNP chunking module to recognise multi-word OOV terms and automatically extract translations through mining the web. In conclusion, a combination of these techniques provides a significant improvement in CLQA effectiveness. We also observed that the syntactic form of a question can be impaired during query translation, and thus potentially degrading the overall CLQA system performance. There are several aspects where we can improve the performance of the system:

- Combine both rule-based pattern matching, statistical methods and learning methods to assign each question to a question type more accurately. Some questions are ambiguous when using only rule-based pattern matching based on keywords.
- Make different and more detailed policies for each question type to find the correct answer. We made only four different kinds of coarse grain polices: (PERSON, LOCATION, ORGANIZATION), (NUMEX), (ARTIFACT), (DATE, TIME). This obviously impairs the performance accuracy of our QA system.
- In Chinese monolingual QA, the accuracy of ICTCLAS probably accounts for much of the deficiency. We need to either improve it or try alternatives in order to correctly identify named entity types.

7 Acknowledgement

This work is supported by a China–Ireland Science and Technology Collaboration Research Fund award under Grant No. CI-2004-12 and China High Technology 973 project under Grant No. 2004CB318109.

This work is also partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD — project MultiMATCH contract IST-033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

References

[Adafre and de Rijke, 2006] Adafre, S. F. and de Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69, Trento, Italy.

[Federico and Bertoldi, 2002] Federico, M. and Bertoldi, N. (2002). Statistical cross-language information retrieval using N-best query translations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 167–174, Tampere, Finland. ACM Press.

[Judge et al., 2005] Judge, J., Guo, Y., Jones, G. J. F., and Wang, B. (2005). An analysis of question processing of english and chinese for the ntcir-5 cross-language question answering task. In *Proceedings of the Fifth NTCIR Workshop on Research in Information Access Technologies*, pages 545–551, Tokyo, Japan.

[Kishida et al., 2005] Kishida, K., hua Chen, K., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., and Myaeng, S. H. (2005). Overview of CLIR task at the fifth NTCIR workshop. In *Proceedings of the 5th NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, Tokyo, Japan. National Institute of Informatics, Japan.

[Kudo and Matsumoto, 2001] Kudo, T. and Matsumoto, Y. (2001). Chunking with support vector machines. In *NAACL '01: 2nd meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

[Kwok, 2000] Kwok, K. L. (2000). Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, pages 173–179, Hong Kong, China. ACM Press.

[Marcus et al., 1994] Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

[Ramshaw and Marcus, 1995] Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In Yarovsky, D. and Church, K., editors, *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey. Association for Computational Linguistics.

[Ratnaparkhi, 1996] Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In Brill, E. and Church, K., editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey.

[Robertson et al., 1995] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and M.Gatford (1995). Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*. NIST.

[Sang and Buchholz, 2000] Sang, E. F. T. K. and Buchholz, S. (2000). Introduction to the CoNLL-2000 shared task: chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*, pages 127–132, Morristown, NJ, USA. Association for Computational Linguistics.

[Sasaki et al., 2007] Sasaki, Y., Lin, C.-J., hua Chen, K., and Chen, H.-H. (2007). Overview of the ntcir-6 cross-lingual question answering (CLQA) task. In *Proceedings of the Sixth NTCIR Workshop on Research in Information Access Technologies*, pages 153–163, Tokyo, Japan.

[Zhang et al., 2007] Zhang, S., Wang, B., and Jones, G. J. F. (2007). ICT-DCU question answering task at ntcir6. In *Proceedings of the Sixth NTCIR Workshop on Research in Information Access Technologies*, pages 154–167, Tokyo, Japan.

[Zhang and Vines, 2004] Zhang, Y. and Vines, P. (2004). Using the web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, pages 162–169, Sheffield, UK. ACM Press.

[Zhang et al., 2005] Zhang, Y., Vines, P., and Zobel, J. (2005). Chinese OOV translation and post-translation query expansion in chinese–english cross-lingual information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):57–77.