Keyword Translation from English to Chinese for Multilingual QA[1]

Frank Lin and Teruko Mitamura

Language Technologies Institute
Carnegie Mellon University
{frank,teruko}@cs.cmu.edu

Abstract. The Keyword Translator is a part of the Question Analyzer module in the JAVELIN Question-Answering system; it translates the keywords that are used to query documents and extract answers. Much work has been done in the area of query translation for CLIR or MLIR, yet most have focused on methods using hard-to-obtain and domain-specific resources, with evaluation often based on retrieval performance rather than translation correctness. In this paper we describe methods combining easily accessible, general-purpose MT systems to improve keyword translation correctness. We also describe methods that utilize the question sentence available to a question-answering system to improve translation correctness. We will show that using multiple MT systems and the question sentence to translate keywords from English to Mandarin Chinese can improve keyword translation correctness.

1 Introduction

Query translation plays an important role in Cross-Language Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR) applications. CLIR and MLIR systems can either translate the query (usually a set of keywords) used to retrieve the documents or translate the documents themselves. Since translating documents is more expensive computationally, most CLIR and MLIR systems chose to do query translation. Similarly, the translation of keywords (words used to retrieve relevant documents and extract answers) in a multilingual question-answering system is crucial when the answer to the question is in a document written in a different language.

The Keyword Translator is not a stand-alone system; it is a part of the Question Analysis module, which is part of the JAVELIN (Justification-based Answer Valuation through Language Interpretation) multilingual open-domain question-answering system [1]. The Question Analysis module is responsible for analyzing (syntactically and semantically) the question sentence, classifying the question, parsing the question, and identifying the keywords within the question sentence. The Keyword Translator, as a sub-module of the Question Analysis module, translates keywords into other languages so other modules can use them to find answers in those languages.

[1] This work was supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program.

Typically, the quality of query translation in CLIR and MLIR is measured by the performance of the CLIR or MLIR system as a whole (precision and recall). However, rather than looking at the overall information retrieval or question-answering performance, in this paper we will focus on the translation correctness of keywords.

Now we will briefly survey previous work relevant to the Keyword Translator.

Query translation, or keyword translation, are typically done using machine-readable dictionaries (MRD), ready-to-use MT systems, parallel corpora, or any combination of the three. Since JAVELIN is an open-domain question-answering system, we do not consider methods that mainly use parallel corpora for query translation. Although comparisons have been made between simple dictionary-based approach and MT-based approach [2] with MT-based having better performance in CLIR, we will consider improved versions to both approaches.

Dictionary-based approaches are popular since the bilingual dictionaries are easily obtained for high-density language pairs; however, dictionary-based methods must overcome the problem of limited coverage and sense disambiguation [3].

Chen et al. [4] described a technique using a combination of dictionaries and search engine to provide adequate coverage, but the results were disappointing due to low coverage. Gao et al. [5] described a method which tries to solve the low-coverage problem by using corpus statistics. They built a noun-phrase translation model and tried to solve the sense-disambiguation problem by using words within the same query as the contextual information and calculating the correlation using the target corpora. Their method showed significant improvement over simple dictionary translation. However, this method requires the building of parallel corpora, and multilingual parallel corpora are difficult to obtain. Jang et al. [6] experimented with sense-disambiguation using a bilingual dictionary and statistics from a collection of documents in the target language with good results, but this method and others [7] that similarly use corpus statistics do not generally solve the low-coverage problem. Seo et al. [8] presented a solution to this problem by using two bilingual dictionaries, a general dictionary and a biographical dictionary (for translating proper names not covered by the general dictionary). Pirkola [9] combined a general dictionary with a domain-specific dictionary, together with structured queries, to achieve CLIR performance almost as good as monolingual IR. However, the methods described by Seo et al. and Pirkola both need a special domain-dictionary, which are not practical in a multilingual open-domain setting.

Generally, MT-based approaches have wider coverage than dictionary-based approaches, since many MT systems available translate common proper names and phrases. However, the quality of translation and sense disambiguation is fully dependent upon the MT system employed. Therefore, people generally do not rely on a single MT system for CLIR translation.

Lam-Adesina and Jones [10] merged results from two MT systems, together with term weighing and query expansion, which improved retrieval performance. Work has also been done in concatenating query translation results from MT systems and dictionaries [11] and in merging documents retrieved by dictionary-translated keywords and documents retrieved by MT-translated keywords [12]. However, since these approaches focus on concatenation of query translations provided by different sources and their performance measure is based on document retrieval, it is difficult to measure in isolation the performance of these approaches in translating queries.

With the above observations, we chose an MT-based approach for the Keyword Translator.

2 The Keyword Translator

The Keyword Translator has two distinguishing features: 1) it uses multiple MT systems and tries to select one correct translation candidate for each keyword and 2) it utilizes the question sentence available to a question-answering system as a context in which to translate the word in the correct sense.

We choose to use multiple MT systems and to utilize the question sentence based on the following assumptions:

- Using more than one MT system gives us a wider range of keyword translation candidates to choose from, and the correct translation is more likely to appear in multiple MT systems than a single MT system, and
- Using the question sentence available to a Question-Answering system gives us a context in which to better select the correct translation candidate.

Based on our assumption, we conducted an experiment to study ways to score translation candidates that would result in correct keyword translations.

3 The Experiment

We study the performance (translation correctness) of the Keyword Translator using three free web-based MT systems with different keyword scoring algorithms, from English to Chinese.

For building the models and tuning the parameters, we compiled a list of 50 English questions by selecting different types of questions from TREC-8, TREC-9, and TREC-10. Then we ran the Question Analyzer module over these questions to get the English keywords. The English question sentences and keywords are the input to the Keyword Translator.

For testing the translator, we randomly selected another set of 50 English questions from the same source, making sure that no questions on the training set appears on this testing set. The Question Analyzer module was also used to produce the keywords for input to the Keyword Translator.

3.1 Evaluation Criteria

The correctness of the translation is evaluated by hand; a translated keyword is considered correct if: 1) it is a valid translation of the original according to the context of the question sentence, and, 2) it is fully translated.

For example, in an MT system "Vesuvius" is not translated as "维苏威," which is the correct translation of "Vesuvius." Instead, "Vesuvius" is returned un-translated. In this case the translation is incorrect. However, there is an exception when an English word's Chinese counterpart is the English word itself. For example, "Photoshop" is a valid Chinese translation of the English word "Photoshop" because Photoshop is never transliterated or translated into Chinese characters in Chinese documents.

The translation correctness will be evaluated based on the percentage of keywords translated correctly.

3.2 MT Systems and Baseline Model

We do keyword translation using three general-purpose MT systems freely available on the internet [2]; we will refer to them as S1 (www.systranbox.com), S2 (www.freetranslation.com), and S3 (www.amikai.com).

We first test the three systems independently on the training set, with these results:

Table 1. Performance of independent MT systems.

	S^1	S^2	S^3
Correct/ Total Keywords	104/125	102/125	92/125
Accuracy	83.20%	81.60%	73.60%

The scoring algorithm of the baseline model is simply to score keywords which appear in two or more MT systems higher than those appear in only one MT system. In the case where all three have the same score (in the baseline model, this happens when all three MT systems give different translations), the translation given by the best-performing individual system, S1, would be selected. And in the case where S2 and S3 tie for the highest scoring keyword, S2 is chosen over S3. This is so that the accuracy of multiple MT's would be at least as good as the best individual MT when there is a tie in the scores. However, with the baseline model, using more MT systems may not give us improvement over the 83.2% accuracy of the best individual system. This is due to the following reasons: 1) For many keywords MT systems all disagree with one another, so S1 is chosen by default, and 2) MT systems may agree on the wrong translation. The results from the baseline model show that there is much room for improvement in the scoring algorithm. For comparing the improvement of multiple MT systems over individual systems, we use S1 for the single MT model, S1 and S2 for the two-MT model, and S1, S2, and S3 for the three-MT model. W choose the MT systems to use for evaluating one-MT and two-MT performance based on Table 1, using the best MT first. This way the improvement made by adding more MT's is not inflated by adding MT's with higher performance.

4 Keyword Scoring Metrics

In this section we describe the metrics used to score or penalize keywords to improve over the baseline model.

[2] Note that reliability issues that come with accessing MT systems remotely can be solved by purchasing and installing them locally.

4.1 Segmented Word-Matching with Partial Word-Matching

Many incorrect translations are incorrect due to sense ambiguity. For sense disambiguation we propose segmented word-matching with partial word-matching. This method is based on our assumption that question sentences provide a “context” in which to translate the keywords in the correct sense.

First we translate the question sentences using the same MT systems used to translate the keywords. Then we segment the sentence (since in written Chinese word boundaries are not marked by spaces) using a combination of forward-most matching and backward-most matching. Forward-most matching (FMM) is a greedy string matching algorithm that starts from left to right and tries to find the longest substring of the string in a word list (we use a word list of 128,455 Mandarin Chinese words compiled from different resources). Backward-most matching (BMM) does the same thing from right to left. We keep the words segmented from FMM and BMM in an array that we call “the segmented sentence.” See Figure 1:

Fig. 1. An example of segmentation.

Sentence: 什么是高尔夫球的直径
FMM: 什么是/高尔夫球/的/直径/
BMM: 什么是/高/尔/夫/球的/直径/
Segmented Sentence: 什么是/夫/尔/球的/的/直径/高/高尔夫球/

After we have the segmented Chinese sentence, we try to match the translated keywords to the segmented sentence, and keywords that match the segmented sentence are scored higher, since they are more likely to be translated in the context of the question sentence.

A feature of the Chinese language is that words sharing the characters are often semantically related. Using this idea, a translated keyword is considered to “partially match” the segmented sentence if the keyword have characters in common with any word in the segmented sentence. A partially matched keyword does not get as high a score as a fully matched word. A fully matched word would score higher than a partially matched word, and a partially matched word would score higher than a word that does not match at all. As we have mentioned previously, if keywords words have the same score, then by default, S1 is selected over S2 or S3, and S2 is selected over S3. When a keyword translation partially matches a word in the segmented sentence, *it is the word in the segmented sentence that is used as the keyword.* Figure 2 shows examples of fully matched and partially matched words:

Fig. 2. Examples of partial word-matching.

非洲 fully matches 非洲	– because these two strings are identical
加利福尼亚 partially matches 加州	– because “加” is a common substring of both words
猎犬 partially matches金黄猎犬	– because “猎犬” is a common substring of both words

Figure 3 shows a full example of segmented word-matching with partial word-matching. We see from the example that “最高的” is chosen over “高” since a fully matched word scores higher than a partially matched word. In both S1 and S2, “山” partially matches “山的,” and “山的” is selected as the translation instead of “山.”

Fig. 3. An example of segmented word-matching with partial word-matching.

English Sentence: What is the name of the highest mountain in Africa?
English Keywords: highest/mountain/name/Africa/
S1: Sentence Translation: 高山的名字是什么在非洲?
 Segmentation: 什么/名字/在/山的/是/的/非洲/高/高山/
 Keyword Translations:
 最高 (highest) partially matches 高
 山 (mountain) partially matches **山的**
 名字 (name) matches **名字**
 非洲 (Africa) matches **非洲**
S2: Sentence Translation: 在非洲的最高的山的名字是什么？
 Segmentation: 什么/名字/在/山的/是/最高的/非洲的/
 Keyword Translations:
 最高的 (highest) matches **最高的**
 山 (mountain) partially matches **山的**
 名字 (name) matches **名字**
 非洲 (Africa) partially matches 非洲的
Final Keywords
最高的/山的/名字/非洲/

4.2 Full Sentence Word-Matching (with or without Fall Back to Partial Word-Matching)

One potential problem with the previous (4.1) algorithm is that the word list we used may not provide adequate coverage to properly segment the question sentence.

In order to solve the problem with the limited coverage of the word list, we also tried word-matching on the entire un-segmented sentence. This is a simple string matching to see if the translated keyword is a substring of the translated question sentence. There are two variations to this metric; in the case where word-matching on the entire un-segmented question sentence fails, we can either fall back to partial word-matching on the segmented sentence or not fall back to partial word-matching. Figures 4 shows an example of full sentence word-matching with fall back to partial word-matching and Figure 5 an example of shows full sentence word-matching without fall back to partial word-matching:

Fig. 4. Full sentence word-matching with fall back to partial word-matching.

English Sentence:
What is the name of the highest mountain in Africa?
English Keywords:
highest/mountain/name/Africa/
Sentence Translation:
高山的名字是什么在非洲?
Segmentation:
什么/名字/在/山的/是/的/非洲/高/高山/
Keyword Translations:
最高 (highest) partially matches 高
山 (mountain) matches question sentence
名字 (name) matches question sentence
非洲 (Africa) matches question sentence

Fig. 5. Full sentence word-matching without fall back.

English Sentence:
What is the name of the highest mountain in Africa?
English Keywords:
highest/mountain/name/Africa/
Sentence Translation:
高山的名字是什么在非洲?
Keyword Translations:
最高 (highest) does not match sentence
山 (mountain) matches sentence
名字 (name) matches sentence
非洲 (Africa) matches sentence

In both Figure 4 and 5, "山," "名字," and "非洲" matches the question sentence but "最高" does not. In Figure 4, partial-matching is used to match "最高" to "高," therefore "高" is considered to be the partial-matched translation and is scored as a partial matched keyword. In Figure 5, "最高" does not go through partial matching so no score is added to the keyword.

4.3 Penalty for Keywords Not Fully Translated

As we explained in section 3.1, keywords that are not fully translated are usually considered incorrect. So we put a penalty on the score of keywords that are not fully translated. This is done by simply checking if [A-Z][a-z] appear in the keyword string. For example, the score for the translation candidate "玛丽亚・Theresa" would be penalized because it is not a full translation. So another translation candidate, 玛利亚·特立莎, would be selected instead.

4.4 Scoring

Each keyword starts with an initial score of 0.0, and as different metrics are applied, numbers are added to or subtracted from the score. Table 2 shows actual numbers used for this experiment; the same scoring scheme is used by all scoring metrics when applicable.

Table 2. Scoring scheme.

A	B	C	D
Full Match	**Partial Match**	**Support by >1 MT**	**Not Fully Translated**
+1.0	+0.5	+0.3	-1.3

The general strategy behind the scoring scheme is as follows: keywords with full match (A) receive the highest score, keywords with partial match (B) receive the second highest score, keywords supported by more than one MT system (C) receive the lowest score, and keywords not fully translated (D) receive a penalty to their score. All of the above can be applied to the same keyword except for A and B, since a keyword cannot be both a full match and a partial match at the same time.

Support by more than one MT system receives the least score because in our experiment it has shown to be the least reliable indication of a correct translation. Full match has shown to be the best indicator of a correct translation, therefore it receives the highest score and it is higher than the combination of partial match and support by more than one MT system. Keywords not fully translated should be penalized heavily, since they are generally considered incorrect; therefore the penalty is set equal to the highest score possible, the combination of A and C. This way, another translation that is a full translation has the opportunity to receive a higher score. With the above general strategy in mind, the numbers were manually tuned to give the best result on the training set.

5 Results

We construct different 7 models (including the baseline model) by combining various scoring metrics. In all models we use the baseline metric that adds score to keyword translation candidates that are supported by more than one MT system. Table 3 shows the abbreviation for each metric, the description of each metric, the section of this paper that describe each metric, and the scoring (refer to Table 2 column headings to look up scoring) that applies to each metric:

Table 3. Abbreviation for scoring metrics and the scoring that is applied

	Description (section)	**Scoring**
B	Baseline (3.3)	C
S	Segmented Word-Matching and Partial Word-Matching (4.1)	A,B
F^1	Full Sentence Word-Matching without Fall Back to Partial Word-Matching (4.2)	A
F^2	Full Sentence Word-Matching with Fall Back to Partial Word-Matching (4.2)	A,B
P	Penalty for Partially Translated or Un-Translated Keywords (4.3)	D

Table 4 shows the percentage of keywords translated correctly using different models on the training set, which consists of 125 keywords from 50 questions. Table 5 shows the improvement of different models over the baseline model based on Table 4:

Table 4. Keyword translation accuracy of different models on the training set.

Model	**S^1**	**S^1S^2**	**$S^1S^2S^3$**
B	83.20%	83.20%	83.20%
B+S	58.40%	64.80%	66.40%
B+F^1	83.20%	85.60%	87.20%
B+F^2	80.80%	84.00%	86.40%
B+P	83.20%	83.20%	83.20%
B+F^1+P	**83.20%**	**89.60%**	**90.40%**
B+F^2+P	80.80%	88.00%	89.60%

Table 5. Improvement of different models over the baseline model on the training set.

Model	**S^1**	**S^1S^2**	**$S^1S^2S^3$**
B+S	-29.81%	-22.12%	-20.19%
B+F^1	0.00%	2.88%	4.81%
B+F^2	-2.88%	0.96%	3.85%
B+P	0.00%	0.00%	0.00%
B+F^1+P	**0.00%**	**7.69%**	**8.65%**
B+F^2+P	-2.88%	5.77%	7.69%

Table 6 shows the percentage of keywords translated correctly using different models on the test set, which consists of 147 keywords from 50 questions. Table 7 shows the improvement over the baseline based on Table 6:

Table 6. Keyword translation accuracy of different models on the test set.

Model	**S^1**	**S^1S^2**	**$S^1S^2S^3$**
B	78.23%	78.23%	78.91%
B+S	59.86%	61.90%	64.63%
B+F^1	78.23%	80.27%	80.95%
B+F^2	76.19%	75.51%	78.91%
B+P	78.23%	78.23%	78.91%
B+F^1+P	**78.23%**	**82.99%**	**85.71%**
B+F^2+P	76.19%	78.23%	83.67%

Table 7. Improvement of different models over the baseline model on the test set.

Model	**S^1**	**S^1S^2**	**$S^1S^2S^3$**
B+S	-23.48%	-20.87%	-18.10%
B+F^1	0.00%	2.61%	2.59%
B+F^2	-2.61%	-3.48%	0.00%
B+P	0.00%	0.00%	0.00%
B+F^1+P	**0.00%**	**6.08%**	**8.62%**
B+F^2+P	-2.61%	0.00%	6.95%

Note that all models which use only one MT system does not improve over the baseline model because no improvement can be made when there is no alternative translations to choose from. However, single-MT models can degrade due to partial word-matching using segmented sentence. We will discuss problems with using segmented sentence and other issues in the next section.

6 Discussions and Conclusion

From the results of different models on the training set and test set, we make the following observations:

1. In almost all models, using additional MT systems do not seem to degrade translation correctness but has the potential to improve translation correctness.
2. As shown in model B+F1, using word-matching on the translated question sentence for sense disambiguation does improve translation correctness.
3. From results of models with S and F2 we see that scoring metrics requiring word list segmentation not only does not improve the translation, they can degrade the translation beyond the baseline model. This method relies on the word list to do the segmentation, and word lists' limited coverage degrades the translation greatly.
4. Full sentence word-matching (F1) with penalty for partially or un-translated keywords (P) yields the best results. Although P does not improve the baseline by itself, it boosts the performance of F1 greatly when combined.

From the above four points and other observations, we briefly describe the pros and cons of using the different scoring metrics in Table 8. The asterisk (*) indicates that this experiment does not validate the statement due to the limited coverage of the word list we used.

Table 8. Pros and cons of the scoring metrics.

	Pros	**Cons**
B	Tie-breaker when two MT systems have the same score	Provides little improvement
S	May work well with an adequate word list for segmentation*	Very poor without adequate word list
F^1	Provides contextual disambiguation; needs no segmentation	Does not do partial matching
F^2	Provides contextual disambiguation	Needs adequate word list
P	Good when individual MT systems lack word coverage	

From Table 8 we can see why model B+F1+P with all three MT systems out performs the others. It 1) uses three MT systems, 2) penalizes keywords that are not fully translated, and 3) does word sense disambiguation without relying on segmentation which needs a word list with adequate coverage, and such a word list may be difficult to obtain. Thus for translating keywords using general MT systems, we can suggest that 1) it is better to use more MT systems if they are available, 2) always penalize un-translated words because different MT systems have different word coverage, and 3) in a setting where resources are limited (small word lists), it is better not to use methods involving segmentation.

In this paper, we first present the general problem of keyword translation in a multilingual open-domain question-answering system. Then based on this general problem, we chose an MT-based approach using multiple free web-based MT systems. And based on our assumption that using multiple MT systems and the question sentence can improve translation correctness, we present several scoring metrics that can be used to build models that choose among keyword translation candidates. Using these models in an experiment, we show that using multiple MT system and using the question sentence to do sense disambiguation can improve the correctness of keyword translation.

References

1. Nyberg, E., Mitamura, T., Callan, J., Carbonell, J., Frederking, R., Collins-Thompson, K., Hiyakumoto, L., Huang, Y., Huttenhower, C., Judy, S., Ko, J., Kupse, A., Lita, L., Pedro, V., Svoboda, D., and Van Durme, B.: The JAVELIN Question-Answering System at TREC 2003: A Multi-Strategy Approach with Dynamic Planning. In: Proceedings of the 12th Text REtrieval Conference (2003).
2. Yang, Y. and Ma, N.: CMU in Cross-Language Information Retrieval at NTCIR-3. In: Proceedings of the Third NTCIR Workshop (2003).
3. Chen, H-H., Bian, G-W. and Lin, W-C.: Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval. In: Proceedings of the Third NTCIR Workshop (2002).
4. Chen, A., Jiang, H. and Gey, F.: Combining Multiple Sources for Short Query Translation in Chinese-English Cross-Language Information Retrieval. In: Proceedings of the Fifth International Workshop Information Retrieval with Asian Languages (2000).
5. Gao, J., Nie, J-Y., Xun, E., Zhang, J., Zhou, M. and Huang, C.: Improving Query Translation for Cross-Language Information Retrieval using Statistical Models. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (2001).
6. Jang, G. M., Kim, P., Jin, Y., Cho, S-H. and Myaeng, S. H.: Simple Query Translation Methods for Korean-English and Korean-Chinese CLIR in NTCIR Experiments. In: Proceedings of the Third NTCIR Workshop (2003).
7. Lin, W-C. and Chen, H-H.: Description of NTU Approach to NTCIR3 Multilingual Information Retrieval. In: Proceedings of the Third NTCIR Workshop (2003).
8. Seo, H-C., Kim, S-B., Kim, B-I., Rim, H-C. and Lee, S-Z.: KUNLP System for NTCIR-3 English-Korean Cross-Language Information Retrieval. In: Proceedings of the Third NTCIR Workshop (2003).
9. Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In: SIGIR 98 (1998).
10. Lam-Adesina, A. M. and Jones, G. J. F.: EXETER AT CLEF 2002: Experiments with Machine Translation for Monolingual and Bilingual Retrieval. In: Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002 (2002).
11. Kwok, K. L.: NTCIR-3 Chinese, Cross Language Retrieval Experiments Using PIRCS. In: Proceedings of the Third NTCIR Workshop (2003).
12. Zhang, J., Sun, L., Qu, W., Du, L., Sun, Y., Fan, Y. and Lin, Z.: ISCAS at NTCIR-3: Monolingual, Bilingual and MultiLingual IR Tasks. In: Proceedings of the Third NTCIR Workshop (2003).