A Low Cost Machine Translation Method for Cross-Lingual Information Retrieval

David B. Bracewell, Fuji Ren, and Shingo Kuroiwa *

Abstract

In one form or another language translation is a necessary part of cross-lingual information retrieval systems. Often times this is accomplished using machine translation systems. However, machine translation systems offer low quality for their high costs. This paper proposes a machine translation method that is low cost while improving translation quality. This is done by utilizing multiple web based translation services to negate the high cost of translation. A best translation is chosen from the candidates using either consensus translation selection or statistical analysis. Which to use is determined by a heuristic rule that takes into account that most web based translation services are of similar quality and that machine translation still produces relatively poor results. By choosing the best translation the method is able to increase translation quality over the base systems, which is verified by the experimentation.

Keywords: Machine Translation, Low Cost, Cross-Lingual Information Retrieval

1 Introduction

Cross-Lingual Information Retrieval (CLIR) is an area of information retrieval that has seen a boom in research in the past few years. The goal of CLIR is to allow users to make queries in one language and retrieve relevant documents in other languages. For example, searching in English for “Natural Language Processing” and retrieving documents in Japanese about “shizen gengo shori” (natural language processing).

Most CLIR systems use some type of translation, whether it be a simple bilingual dictionary or a full machine translation (MT) system. Which type of translation to use is typically decided by what the system will translate. CLIR systems typically translate either queries or documents.

Query translation is often accomplished using bilingual dictionaries, such as [5] and [11]. The main problem with this approach is the difficulty in creating the dictionaries, especially in creating ones that rival the size of the ones used within machine translation systems. Moreover, such techniques are generally only used for translating query words and cannot effectively handle short phrases.

Machine translation systems are often used for document translation. Researchers have found that in certain instances document translation performs better than query translation [12] and [6]. In addition, machine translation systems can be used for translating queries, summaries, etc. As such, using a machine translation system makes more sense for the general case of CLIR.

Machine translation has been widely studied since its inception in the 1950s. Research has been done on translating between a wide number of languages, such as Japanese-English [17], Chinese-English [13] and German-French [20]. There are three main factors that hinder the use of machine translation: speed, cost and translation quality.

Carbonell et al. point out that machine translation is computationally expensive and in some cases impractical to use [4]. However, with the increases in computing power and the development of faster algorithms this is becoming less of a problem.

The cost of machine translation can be attributed to either the monetary cost or creation cost. The monetary cost covers the licensing fees needed to use a preexisting system and the purchase of dictionaries, corpora, etc. Creation cost includes the cost of creating bilingual dictionaries, parallel corpora and the actual construction and evaluation of the MT system. One possibility to get around the monetary cost and creation cost is to use web based translation services.

The final factor in using machine translation is the quality of translation. Even with its long history and extensive research, the quality of translation is still not generally good enough to be acceptable by users as [16] and [15] point out. One way around this is to limit what is translated.

In this paper, a system is proposed that helps alleviate the cost of translation and improve the quality of translation. To do this the output of multiple web based trans-

*Department of Information Science and Intelligent Systems The University of Tokushima, JAPAN Email: {davidb,ren,kuroiwa}@is.tokushima-u.ac.jp Fuji Ren is also with the School of Information Engineering, Beijing University of Posts and Telecommunications Beijing 100876, China

(Advance online publication: 19 February 2008)

lation services are used as candidate translations. The best candidate translation is chosen and returned. Using web based services alleviates the cost of translation and choosing the best translation from candidates, as will be shown, helps improve the quality of translation.

This paper will continue as follows. First, in section 2 some more background on Japanese-English machine translation and ways of combining different machine translation systems will be given. Next, in section 3 an overview of the proposed system will be given. Then, section 4 gives an evaluation of the system. Finally, section 5 discusses future work and gives concluding remarks.

2 Background

Many different approaches have been taken in translating between Japanese and English. [15] used the Super Function, which is a mix between example and pattern based translation. [17] created an example based MT system that translated certain sentence types from Japanese to English. [21] introduced a syntax based translation model that they found gave better results than IBM Model 5. The main problems with these preexisting systems are availability and cost. Many of these systems are still in the research stage and are not available as commercial systems. To recreate their work would require a significant creation cost.

Research has also been done on looking at ways of improving current commercially available systems. Typically, these approaches examine the output of many different systems to determine what the best translation would be. Two ways of doing this are n-gram language models [3] and consensus translations [10].

Burch built a language model using a web crawler and then calculated the probability of each translated sentence using the model [3]. Akiba et al. proposed two methods for selecting better translations from multiple Japanese-English MT systems [1]. The first method uses a combination of a language model and translation model and the second method uses conditional probabilities predicted by a regression tree. The methods, such as the ones just mentioned, typically require not only the cost of the base MT systems, but also the cost of creating a large enough corpus to train a language model from. Moreover, a language model will need to be constructed for each target language the user wishes to use, i.e. for Japanese to English an English language model is needed and for English to Japanese a Japanese language model is needed.

Bangalore et al. used an edit distance alignment based approach to determine the similarity amongst candidate translations [2]. However, alignment can be computationally expensive and has problems with translations that have dramatically different word orders. Matusov et al. proposed determining a consensus translation using a pairwise alignment method to create a confusion network [10]. Pairwise alignment is able to overcome some of the shortcomings of edit distance alignment, but still suffers from possibly expensive string alignment.

3 Proposed System

Previous research on choosing a best translation, using either a consensus translation or a language model, was solely motivated on improving translation quality. As the research was focused on the machine translation problem the goal was increased quality with little regard to computational or creation cost. Perfect machine translation is not the goal, and probably not needed, for cross-lingual information retrieval. Instead the goal should be on cheap and fast translation with relatively good quality. As such, this paper proposes a simple and quick way of deciding the best translation from a set of candidates.

The proposed system uses the outcome of multiple web based machine translation systems and determines the best translation from them. The best translation is either chosen using statistical analysis or by selecting the consensus translation. Using web based translation systems eliminates the purchase and creation cost of machine translation and determining the best translation improves the quality.

The systems makes two heuristic assumptions, shown below.

- The translations from the base MT systems are about equal in quality.
- Machine translation is poor in quality.

The first assumption assumes that the translation systems given to it are about equal in their quality of translation, but differ in actual translations. The difference in translations comes about through the different methods, dictionaries, corpora, etc used in their construction. Using this heuristic, the system tries to choose a consensus translation first. This assumption implies that determining the consensus translation among the candidates will lead to a good translation.

However, sometimes the translations will be the same amongst a number of systems. Typically, this type of consensus is good and desired, but if the translation that is agreed upon is poor in quality then the resulting chosen translation will also be poor in quality. Using the second heuristic assumption we can say that the quality of the consensus translation in such a case may not be the best translation. In order to deal with this situation statistical analysis is performed to determine the best translation.

Combining these heuristics gives the following rule: *If*

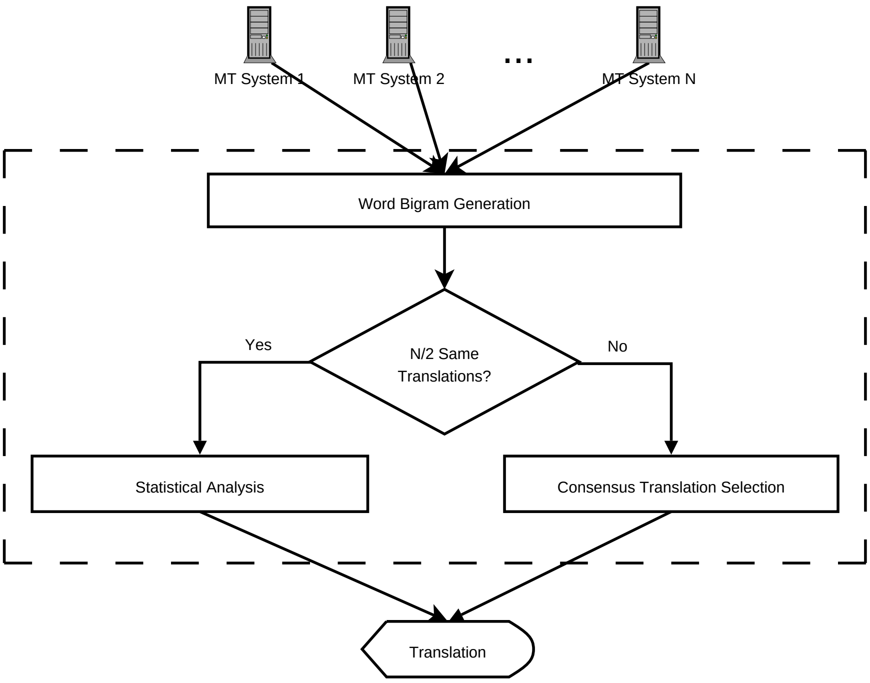


Figure 1: Overview of Proposed Method

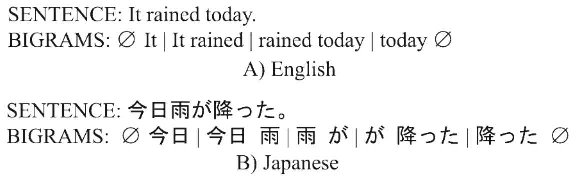
many of the machine translation systems have the same translation then use statistical analysis to choose the best translation otherwise choose the consensus translation. In this paper, "many of the machine translation systems" is defined to be at least half.

An overview of the system can be seen in figure 1. It its divided into three modules: word bigram generation, consensus translation selection and statistical analysis. The choice of whether to perform statistical analysis or determine a consensus translation is governed by the heuristics inspired rule. The following subsections will examine each of the modules in more detail.

3.1 Word Bigram Generation

The system judges and chooses the best translation either by determining the consensus translation or through statistical analysis. To do this word level bigrams are used. Word bigrams allow for quick processing and bigrams in general good give results.

The output of the base translation systems are taken as candidate translations. For each of the candidate translations bigrams are extracted. For English, extracting word level bigrams is as simple as separating the text on whitespace. However, Japanese does not use whitespace in between its words. Because of this, word segmentation needs to be performed before the bigrams are created. Currently, Chasen [9] is used. Figure 2 gives examples of bigram generation for a simple English and Japanese sentence (∅ represents the beginning or end of sentence).


SENTENCE: It rained today.
BIGRAMS: ∅ It | It rained | rained today | today ∅
A) English

SENTENCE: 今日雨が降った。
BIGRAMS: ∅ 今日 | 今日 雨 | 雨 が | が 降った | 降った ∅
B) Japanese

Figure 2: Example Word Bigrams

3.2 Consensus Translation Selection

A consensus translation is created by combining multiple possible translations. Typically, costly string alignment is used to create a confusion network and then the consensus translation is computed by finding the voting on the best path in the network, such as [10]. The advantage is that a new translation that was not present in the base MT systems can be created and the disadvantage is speed.

In contrast, this paper defines a consensus translation as the translation that is most similar to the others. In other words, if the candidate translations are thought about as being a cluster then the consensus translation would be the cluster's centroid. Using this definition, the candidate translation with the shortest total distance between itself and the other candidates would be the consensus. This type of calculation improves on speed, but is not able to create a new translation not seen in the base MT systems.

To compute the distance between candidate translations, their word level bigram representation is used. Equation

1, shows the distance equation. In the equation T_{t_i} is the set of bigrams making up translation t_i. The total distance is the sum of one minus the Dice's coefficient [19]. Dice's coefficient is commonly used to calculate the similarity between words, sentences, etc. using n-grams. In this case it has been transformed into a dissimilarity measure, which makes it ideal for estimating distance.

$$TotalDistance(T_{t_i}) = \sum_{j=1}^{N} \left(1.0 - \frac{2 \times |T_{t_i} \cap T_{t_j}|}{|T_{t_i}| + |T_{t_j}|} \right) \quad (1)$$

The candidate translation with the shortest total distance is then chosen as the consensus. If there is a tie then the winner is determined by the sort order. In the future this can be changed to select the base machine translation system that was previously chosen as being better by the user or system designer.

3.3 Statistical Analysis

When too many of the candidate translations are the same, statistical analysis is performed to choose the best translation. In this paper, statistical analysis means computing the probability that each candidate translation t_i is in the target language $\acute{\ell}$. Typically, as in [3] this is done using a corpus. However, the corpus may or may not be able to cover the language properly. In order to better cover the target language and to eliminate the cost of creating the corpus, this paper uses the web to estimate the target language model.

To do this, the number of hits, $H(x)$, that a bigram has and the number of hits a known stop word in the target language has is used to estimate the probability that the bigram is in the target language. Equation 2 shows how to calculate the probability. In the equation $H(\alpha)$ is the number of hits for a known stop word in $\acute{\ell}$, for example "the" in English.

$$P\left(T_{t_i} = \{b_1, b_2, \cdots, b_n\} | \acute{\ell}\right) = \prod_{j=1}^{n} \frac{H(b_j)}{H(\alpha)} \quad (2)$$

This method alleviates the cost of creating a corpus for the target language. It also may yield better results as the number of words in the pages the search engine covers is very likely to be much larger than any corpus. Currently, the proposed system uses Google to determine $H(x)$.

After a probability is determined for each candidate, the candidates are sorted. The candidate with the highest probability is chosen as the best translation. As with consensus translation selection, in instances where there is a tie, the sort order decides the best translation.

4 Evaluation

The system currently uses three online MT systems as base systems: Google Translation [1], Excite Japan [2] and Babelfish [3]. Bablefish uses technology by Systran [18] for its translation service. Systran uses a single engine to convert between a wide variety of languages and uses NLP technologies like part-of-speech tagging, word segmentation, and semantic domain recognition [18]. Excite Japan is powered by BizLingo [8]. BizLingo is a product from Fujitsu that is a web based solution for Japanese and English translation. Google Translation uses an example based approach and they collect statistics from very large parallel corpora.

Evaluation was done by using a set of 50 Japanese-English bilingual sentences collected from EDP made available by Eijiro [7]. Each sentence was translated by each of the three base translation systems and the proposed system. Two different types of evaluations where then carried out: human evaluation and vocabulary overlap.

4.1 Human Evaluation

In the first evaluation, an independent judge was presented with the source sentence and the four translated sentences. The judge was then asked to choose the best translation. All systems with the best translation were given credit for a correct answer. Table 1 shows the results for translation from English to Japanese and table 2 shows the results for translating from Japanese to English.

Table 1: Human Evaluation Results for English to Japanese

System	Percentage Best
Babelfish	26%
Excite	78%
Google	21%
Proposed	83%

The results show that translating from English to Japanese gave better results than Japanese to English. Among the three base systems, Excite was the best. Babelfish performed better than Google in translating English to Japanese, but Google edged Babelfish out in Japanese to English. The proposed system was able to achieve a 5% increase over Excite in English to Japanese translation and a 2% increase over Excite in Japanese to English translation.

[1] http://translate.google.com
[2] http://www.excite.co.jp/world/english/
[3] http://world.altavista.com/tr

Table 2: Human Evaluation Results for Japanese to English

System	Percentage Best
Babelfish	17%
Excite	83%
Google	19%
Proposed	85%

The results do not tell how good the quality of translation was only that the proposed system was able to accurately choose the best translation available. With that said the following informal observations were made. Overall, the translations were very poor, but good enough for CLIR. Translating from English to Japanese yielded better quality translations than going from Japanese to English.

4.2 Vocabulary Overlap

The second evaluation looked at the vocabulary overlap between the translated sentences and the given translation in the bitext. Measuring the vocabulary overlap between sentences has been used to judge their similarity for different tasks, such as summarization [14]. Here, the assumption is made that the more overlap there is between the machine translated text and the true translation the more the information given in both is the same.

The overlap coefficient [19] was used to compute the overlap between the machine translated sentence and the original human translation. Table 3 shows the results for English to Japanese translation and table 4 shows results for Japanese to English translation.

Table 3: Vocabulary Overlap Results for English to Japanese

System	Micro Averaged Overlap
Babelfish	27%
Excite	27%
Google	27%
Proposed	39%

Table 4: Vocabulary Overlap Results for Japanese to English

System	Micro Averaged Overlap
Babelfish	22%
Excite	23%
Google	22%
Proposed	31%

These results show the micro averaged overlap for the sentences. The higher the overlap the more the vocabulary is the same. It does not, however, take word order into account. The results show that the proposed system does well in picking sentences that have a higher overlap in vocabulary with real human translations. The other three systems have a similar or same overlap. While, informally, the translations were of poor quality the overlap between the proposed system's translations and human translations is encouraging for CLIR purposes.

5 Conclusion and Future Work

This paper presented a low cost machine translation system that is useful for cross-lingual information retrieval. By using web based machine translation services, the system alleviates the monetary and creation cost often associated with machine translation systems. By examining the output of multiple systems and choosing the best one from the candidates it was also shown to be able to improve overall translation quality.

Determining the best translation is either done by choosing the consensus translation or using statistical analysis. Which approach is used is determined by the translations that are given from the base systems. When less than half of the translations are the same, the consensus translation is chosen as the best. When half or more of the translations are the same the system performs statistical analysis by estimating the probability that the sentence is in the language using Google as a language model. The system was able give a better overall quality of translation than just using one of the MT systems alone. While the quality of translation probably is not close to the start-of-the-art systems, it is adequate enough for the purposes of CLIR.

In the future, we will look at adding more translation services to the system. In addition, we would like to examine the use of the system in an actual CLIR environment, beyond the query, summary and phrase translation we currently use it for. Also, as we increase the languages used in our CLIR system to beyond just Japanese and English, we will look at adding extra services and if needed using pivot languages for translation.

Acknowledgment

This research has been partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid for Scientific Research (B), 14380166, 17300065, Exploratory Research, 17656128.

References

[1] Yasuhiro Akiba, Taro Watanabe, and Eiichiro Sumita. Using language and translation models to select the best among outputs from multiple mt

systems. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[2] B. Bangalore, G. Bordel, and G. Riccardi. Computing consensus translation from multiple machine translation systems. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 351–354, 2001.

[3] C. Callison-Burch. Upping the ante for best of breed machine translation providers. In *Proceedings of ASLIB Translating and the Computer*, 2001.

[4] J.G. Carbonell, Y. Yang, R.E. Frederking, R.D. Brown, Yibing Geng, and Danny Lee. Translingual information retrieval : A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 708–715, 1997.

[5] A. Chen. Phrasal translation for english-chinese cross language information retrieval. In *Proceedings of the 2000 International Conference on Chinese Language Computing*, 2000.

[6] Aitao Chen and Fredric C. Gey. Combining query translation and document translation in cross-language retrieval. In *In Cross-Language Evaluation Forum: Working Notes for the CLEF 2003 Workshop*, 2003.

[7] Eijiro. Edp. [Online], http://www.eijiro.jp/.

[8] Fujitsu. Accela bizlingo. [Online], http://www.fujitsu.com.

[9] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. Morphological analysis system chasen version 2.2.9 manual. Technical report, Nara Institute of Science and Technology, 2002.

[10] E. Matusov, N. Ueffing, and H. Ney. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *In Proceedings of EACL 2006*, 2006.

[11] Craig J. A. McEwan, Iadh Ounis, and Ian Ruthven. Building bilingual dictionaries from parallel web documents. In *ECIR*, pages 303–323, 2002.

[12] Douglas W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 472–483, London, UK, 1998. Springer-Verlag.

[13] Martha Palmer and Zhibiao Wu2. Verb semantics for english-chinese translation. *Machine Translation*, 10:59–92, 1995.

[14] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Inf. Process. Manage.*, 33(2):193–207, 1997.

[15] M. Sasayama, F. Ren, and S. Kuroiwa. Super-function based japanese-english machine translation. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 2003.

[16] Jonathan Slocum. A survey of machine translation: its history, current status, and future prospects. *Comput. Linguist.*, 11(1):1–17, 1985.

[17] Eiichiro Sumita and Hitoshi Iida. Experiments and prospects of example-based machine translation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991.

[18] Systran. Systran language translation technology. [Online], http://www.systransoft.com.

[19] R. C. J. van Rijsbergen. *Information Retrieval: Second Edition.* Butterworth-Heinemann, 1979.

[20] Bernard Vauquois and Christian Boitet. Automated translation at grenoble university. *Comput. Linguist.*, 11(1):28–36, 1985.

[21] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530, Morristown, NJ, USA, 2001. Association for Computational Linguistics.