Adaptation Using Out-of-Domain Corpus within EBMT

Takao Doi, Eiichiro Sumita, Hirofumi Yamamoto
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Kansai Science City, Kyoto, 619-0288 Japan
{takao.doi, eiichiro.sumita, hirofumi.yamamoto}@atr.co.jp

Abstract

In order to boost the translation quality of EBMT based on a small-sized bilingual corpus, we use an out-of-domain bilingual corpus and, in addition, the language model of an in-domain monolingual corpus. We conducted experiments with an EBMT system. The two evaluation measures of the BLEU score and the NIST score demonstrated the effect of using an out-of-domain bilingual corpus and the possibility of using the language model.

1 Introduction

Example-Based Machine Translation (EBMT) is adaptable to new domains. If you simply prepare a bilingual corpus of a new domain, you'll get a translation system for the domain. However, if only a small-sized corpus is available, low translation quality is obtained. We explored methods to boost translation quality based on a small-sized bilingual corpus in the domain. Among these methods, we use an out-of-domain bilingual corpus and, in addition, the language model (LM) of an in-domain monolingual corpus. For accuracy of the LM, a larger training set is better. The training set is a target language corpus, which can be more easily prepared than a bilingual corpus.

In prior works, statistical machine translation (Brown, 1993) used not only LM but also translation models. However, making a translation model requires a bilingual corpus. On the other hand, in some studies on multiple-translation selection, the LM of the target language is used to calculate translation scores (Kaki, 1999; Callison-Burch, 2001). For adaptation, we use the LM of an in-domain target language.

In the following sections, we describe the methods using an out-of-domain bilingual corpus and an in-domain monolingual corpus. Moreover, we report on our experiments.

2 Adaptation Methods

EBMT (Nagao, 1984) retrieves the translation examples that are most similar to an input expression and adjusts the examples to obtain the translation. The EBMT system in our approach retrieves not only in-domain examples, but also out-of-domain examples. When using out-of-domain examples, suitability to the target domain is considered. We tried the following three types of adaptation methods.

(1) Merging equally

An in-domain corpus and an out-of-domain corpus are simply merged and used without distinction.

(2) Merging with preference for in-domain corpus

An in-domain corpus and an out-of-domain corpus are merged. However, when multiple examples with the same similarity are retrieved, the in-domain examples are used.

(3) Using LM

Beforehand, we make an LM of an in-domain target language corpus and, according to the LM, assign a probability to the target sentence of each out-of-domain example.

In the example retrieval phase of the EBMT system, two types of examples are handled differently.

(3-1) From in-domain examples, the most similar examples are retrieved.

(3-2) From out-of-domain examples, not only the most similar examples but also other examples that are nearly as similar are retrieved. In the retrieved examples, examples with the highest probabilities of their target sentences by the LM are selected.

(3-3) From the results of both (3-1) and (3-2), the most similar examples are selected. Examples of (3-1) are used when the similarities are equal to each other.

3 Translation Experiments

3.1 Conditions

In order to evaluate the adaptability of an EBMT with out-of-domain examples, we applied the methods described in Section 2 to the EBMT and evaluated the translation quality in Japanese-to-English translation. We used an EBMT, DP-match Driven transDucer (D^3, Sumita, 2001) as a test bed.

We used two Japanese-and-English bilingual corpora. In this experiment on adaptation, as an out-of-domain corpus, we used Basic Travel Expression Corpus (BTEC, described as BE-corpus in Takezawa, 2002); as an in-domain corpus, we used a telephone conversation corpus (TEL). The statistics of the corpora are shown in Table 1. TEL is split into two parts: a test set of 1,653 sentence pairs and a training set of 9,918. Perplexities reveal the large difference between the in-domain and out-of-domain corpora.

Table 1. Corpus Statistics

	BTEC		TEL	
	Japanese	English	Japanese	English
# of sentences	152,172		11,571	
# of words	1,045,694	909,270	103,860	92,749
Vocabulary size	19,999	12,268	5,242	4,086
Average sentence length	6.87	5.98	8.98	8.02
Perplexity (word trigram)	24.19	28.85	37.22	40.04
	TEL language model		BTEC language model	
	190.77	142.04	57.27	81.26

The translation qualities were evaluated by the BLEU score (Papineni, 2001) and the NIST score (Doddington, 2002). The evaluation methods compare the system output translation with a set of reference translations of the same source text by finding sequences of words in the reference translations that match those in the system output translation. We used the English sentence corresponding to each input Japanese sentence in the test set as the reference translation. Therefore, achieving a better score by the evaluation means that the translation results can be regarded as more adequate translations for the domain.

In order to simulate incremental expansion of an in-domain bilingual corpus and to observe the relationship between corpus size and translation quality, translations were performed with some subsets of the training corpus. The numbers of the sentence pairs are 0, 1000, .. , 5000 and 9918, adding randomly selected examples from the training set.

The LM of the domain's target language was the word trigram model of the English sentences of the training set of TEL. We tried two patterns of training set quantities in making the LM: 1) all of the training set, and 2) the part of the set used for translation examples according to the numbers mentioned above.

3.2 Results

Table 2 shows the BLEU scores from the translation experiment, which show certain tendencies. Generally, by using more in-domain examples, the translation results steadily achieve better scores. The score when using 4,000 in-domain examples exceeded that when using 152,172 out-of-domain examples. Equal merging outperformed using only out-of-domain examples. Merging with in-domain preference outperformed equal merging, and using LM outperformed merging with in-domain preference. Comparing the two cases using LM, using LM made from all of the training set got a slightly better scores than the other, which implies that better LM is made from a larger corpus. All of the adaptation methods are more effective when a smaller-sized in-domain corpus is available. When using no in-domain examples, the effect of using LM made from the entire training set was relatively large.

Table 3 shows the NIST scores for the same experiment. We can observe the same tendencies as in the table of BLEU scores, except that the advantage of using LM made from all of the training set over that from a partial set was not observed.

4 Conclusion and Future Work

A corpus-based approach is able to quickly build a machine translation system for a new domain if a bilingual corpus of that domain is available. However, if only a small-sized corpus is available, a low translation quality is obtained. In order to boost the performance, several methods using out-of-domain data were explored in this paper. The experimental results showed the effect of using an out-of-domain corpus by two evaluation measures, i.e., the BLEU score and the NIST score.

We also showed the possibility of increasing the translation quality by using the LM of the domain's target language. However, the gains from using the LM in the evaluation scores were not significant. We must continue experiments with other corpora and under various conditions. In addition, though we've implicitly assumed a high-quality in-domain corpus, next we'd like to investigate using a low-quality corpus.

Table 2. Experimental results of translation by BLEU scores

# of in-domain examples	0	1,000	2,000	3,000	4,000	5,000	9,918
Using in-domain examples	---	0.0190	0.0602	0.0942	0.1200	0.1436	0.2100
Using out-of-domain examples	0.1099						
Merging equally	0.1099	0.1271	0.1430	0.1590	0.1727	0.1868	0.2303
Merging with preference for in-domain		0.1296	0.1469	0.1632	0.1776	0.1922	0.2333
Using LM of partial training set		0.1361	0.1538	0.1686	0.1829	0.1976	0.2387
Using LM of all training set	0.1225	0.1393	0.1557	0.1716	0.1852	0.1987	

Table 3. Experimental results of translation by NIST scores

# of in-domain examples	0	1,000	2,000	3,000	4,000	5,000	9,918
Using in-domain examples	---	0.0037	0.1130	0.4168	0.7567	1.1619	2.7400
Using out-of-domain examples	1.1126						
Merging equally	1.1126	1.4283	1.7367	2.0690	2.3405	2.6142	3.5772
Merging with preference for in-domain		1.4580	1.7975	2.1343	2.4045	2.7088	3.6255
Using LM of partial training set		1.7454	2.0449	2.3639	2.5825	2.9304	3.7544
Using LM of all training set	1.4404	1.7007	2.0125	2.3484	2.5992	2.8973	

Acknowledgements

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, “A study of speech dialogue translation technology based on a large corpus”.

References

Takezawa, T. et al. 2002. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, Proc. of LREC-2002

Papineni, K. et al. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation, RC22176, September 17, 2001, Computer Science

Doddington, G. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Proc. of the HLT 2002 Conference

Nagao, M. 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in Artificial and Human Intelligence, Elithorn, A. and Banerji, R. (eds.). North-Holland

Sumita, E. 2001 Example-based machine translation using DP-matching between word sequences, Proc. of DDMT Workshop of 39th ACL

Brown, P. F. et al. 1993. The mathematics of statistical machine translation: Parameter estimation, Computational Linguistics, 19(2)

Kaki, S. et al. 1999. Scoring multiple translations using character N-gram, Proc. of NLPRS-99

Callison-Burch, C. et al. 2001. A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines, Proc. of MT Summit VIII