Exploiting Variant Corpora for Machine Translation

Michael Paul[†‡] and **Eiichiro Sumita**[†‡]
† National Institute of Information and Communications Technology
‡ ATR Spoken Language Communication Research Labs
Hikaridai 2-2-2, Keihanna Science City, 619-0288 Kyoto
{Michael.Paul,Eiichiro.Sumita}@{nict.go.jp,atr.jp}

Abstract

This paper proposes the usage of *variant corpora*, i.e., parallel text corpora that are equal in meaning but use different ways to express content, in order to improve corpus-based machine translation. The usage of multiple training corpora of the same content with different sources results in variant models that focus on specific linguistic phenomena covered by the respective corpus. The proposed method applies each variant model separately resulting in multiple translation hypotheses which are selectively combined according to statistical models. The proposed method outperforms the conventional approach of merging all variants by reducing translation ambiguities and exploiting the strengths of each variant model.

1 Introduction

Corpus-based approaches to machine translation (MT) have achieved much progress over the last decades. Despite a high performance on average, these approaches can often produce translations with severe errors. Input sentences featuring linguistic phenomena that are not sufficiently covered by the utilized models cannot be translated accurately.

This paper proposes to use multiple *variant corpora*, i.e., parallel text corpora that are equal in meaning, but use different vocabulary and grammatical constructions in order to express the same content. Using training corpora of the same content with different sources result in translation models that focus on specific linguistic phenomena, thus reducing translation ambiguities compared to models trained on a larger corpus obtained by merging all variant corpora. The proposed method applies each variant model separately to an input sentence resulting in multiple translation hypotheses. The best translation is selected according to statistical models. We show that the combination of variant translation models is effective and outperforms not only all single variant models, but also is superior to translation models trained on the union of all variant corpora.

In addition, we extend the proposed method to multi-engine MT. Combining multiple MT engines can boost the system performance further by exploiting the strengths of each MT engine. For each variant, all MT engines are trained on the same corpus and used in parallel to translate the input. We first select the best translation hypotheses created by all MT engines trained on the same variant and then verify the translation quality of the translation hypotheses selected for each variant.

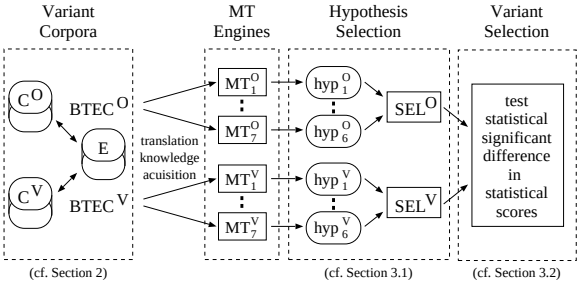

Figure 1: System outline

The outline of the proposed system is given in Figure 1. For the experiments described in this paper we are using two variants of a parallel text corpus for Chinese (C) and English (E) from the travel domain (cf. Section 2). These variant corpora are used to acquire the translation knowledge for seven corpus-based MT engines. The method to select the best translation hypotheses of MT engines trained on the same variant is described in Section 3.1. Finally, the selected translations of different variants are combined according to a statistical significance test as described in Section 3.2. The effectiveness of the proposed method is verified in Section 4 for

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 113–116,
New York, June 2006.

the Chinese-English translation task of last year's IWSLT[1] evaluation campaign.

2 Variant Corpora

The *Basic Travel Expressions Corpus* (BTEC) is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country and cover utterances in travel situations (Kikui et al., 2003). The original Japanese-English corpus consists of 500K of aligned sentence pairs whereby the Japanese sentences were also translated into Chinese.

In addition, parts of the original English corpus were translated separately into Chinese resulting in a variant corpus comprising 162K CE sentence pairs. Details of both, the original (*BTEC*^{O}) and the variant (*BTEC*^{V}) corpus, are given in Table 1, where *word token* refers to the number of words in the corpus and *word type* refers to the vocabulary size.

Table 1: Statistics of variant corpora

corpus	lang	sentence count		avg	word	word
		total	unique	len	tokens	types
BTEC^O	C	501,809	299,347	6.8	3,436,750	40,645
	E	501,809	344,134	8.3	4,524,703	21,832
BTEC^V	C	162,320	97,512	7.1	1,302,761	14,222
	E	162,320	96,988	7.5	1,367,981	9,795

Only 4.8% of the sentences occured in both corpora and only 68.1% of the BTEC^V vocabulary was covered in the BTEC^O corpus.

The comparison of both corpora revealed further that each variant closely reflects the linguistic structure of the source language which was used to produce the Chinese translations of the respective data sets. The differences between the BTEC^O and BTEC^V variants can be categorized into:
(1) **literalness:** BTEC^O sentences are translated on the basis of their meaning and context resulting in freer translations compared to the BTEC^V sentences which are translated more literally;
(2) **syntax:** The degree of literalness also has an impact on the syntactic structure like word order variations (C^V sentences reflect closely the word order of the corresponding English sentences) or the sentence type (*question* vs. *imperative*);
(3) **lexical choice:** Alternations in lexical choice also contribute largely to variations between the corpora. Moreover, most of the pronouns found in the English sentences are translated explicitly in the C^V sentences, but are omitted in C^O;
(4) **orthography:** Orthographic differences especially for proper nouns (*Kanji* vs. *transliteration*) and numbers (*numerals* vs. *spelling-out*).

[1] http://penance.is.cs.cmu.edu/iwslt2005

3 Corpus-based Machine Translation

The differences in variant corpora directly effect the translation quality of corpus-based MT approaches. Simply merging variant corpora for training increases the coverage of linguistic phenomena by the obtained translation model. However, due to an increase in translation ambiguities, more erroneous translations might be generated.

In contrast, the proposed method trains separately MT engines on each variant focusing on linguistic phenomena covered in the respective corpus. If specific linguistic phenomena are not covered by a variant corpus, the translation quality of the respective output is expected to be significantly lower.

Therefore, we first judge the translation quality of all translation hypotheses created by MT engines trained on the same variant corpus by testing statistical significant differences in the statistical scores (cf. Section 3.1). Next, we compare the outcomes of the statistical significance test between the translation hypotheses selected for each variant in order to identify the variant that fits best the given input sentence (cf. Section 3.2).

3.1 Hypothesis Selection

In order to select the best translation among outputs generated by multiple MT systems, we employ an SMT-based method that scores MT outputs by using multiple language (LM) and translation model (TM) pairs trained on different subsets of the training data. It uses a statistical test to check whether the obtained TM·LM scores of one MT output are significantly higher than those of another MT output (Akiba et al., 2002). Given an input sentence, m translation hypotheses are produced by the element MT engines, whereby n different TM·LM scores are assigned to each hypothesis. In order to check whether the highest scored hypothesis is significantly better then the other MT outputs, a multiple comparison test based on the Kruskal-Wallis test is used. If one of the MT outputs is significantly better, this output is selected.

Otherwise, the output of the MT engine that performs best on a develop set is selected.

3.2 Variant Selection

In order to judge which variant should be selected for the translation of a given input sentence, the outcomes of the statistical significance test carried out during the hypothesis selection are employed.

The hypothesis selection method is applied for each variant separately, i.e., the BTEC^O corpus is used to train multiple statistical model pairs (SEL^O) and the best translation (MT$^O_{SEL}$) of the set of translation hypotheses created by the MT engines trained on the BTEC^O corpus is selected. Accordingly, the SEL^V models are trained on the BTEC^V corpus and applied to select the best translation (MT$^V_{SEL}$) of the MT outputs trained on the BTEC^V corpus. In addition, the SEL^O models were used in order to verify whether a significant difference can be found for the translation hypothesis MT$^V_{SEL}$, and, vice versa, the SEL^V models were applied to MT$^O_{SEL}$.

The outcomes of the statistical significance tests are then compared. If a significant difference between the statistical scores based on one variant, but not for the other variant is obtained, the significantly better hypothesis is selected as the output. However, if a significant difference could be found for both or none of the variants, the translation hypothesis produced by the MT engine that performs best on a develop set is selected.

4 Experiments

The effectiveness of the proposed method is verified for the CE translation task (500 sentences) of last year's IWSLT evaluation campaign. For the experiments, we used the four *statistical* (SMT) and three *example-based* (EBMT) MT engines described in detail in (Paul et al., 2005).

For evaluation, we used the BLEU metrics, which calculates the geometric mean of n-gram precision for the MT outputs found in reference translations (Papineni et al., 2002). Higher BLEU scores indicate better translations.

4.1 Performance of Element MT Engines

Table 2 summarizes the results of all element MT engines trained on the BTEC^O and BTEC^V corpora. The result show that the SMT engines outperform the EBMT engines whereby the best performing system is marked with bold-face.

Table 2: BLEU evaluation of element MT engines

SMT	BTEC^O	BTEC^V
MT$_1$	0.4020	0.4633
MT$_2$	0.4474	0.4595
MT$_3$	**0.5342**	**0.5110**
MT$_4$	0.3575	0.4460

EBMT	BTEC^O	BTEC^V
MT$_5$	0.2908	0.3445
MT$_6$	0.2988	0.4100
MT$_7$	0.0002	0.0074

However, depending on the variant corpus used to train the MT engines, quite different system performances are achieved. Most of the element MT engines perform better when trained on the smaller BTEC^V corpus indicating that the given test set is not covered well by the BTEC^O corpus.

4.2 Effects of Hypothesis Selection

The performance of the hypothesis selection method (SEL) is summarized in Table 3 whereby the obtained gain relative to the best element MT engine is given in parentheses. In addition, we performed an "oracle" translation experiment in order to investigate in an upper boundary for the method. Each input sentence was translated by all element MT engines and the translation hypothesis with the lowest word error rate[2] relative to the reference translations was output as the translation, i.e., the ORACLE system simulates an optimal selection method according to an objective evaluation criterion.

Table 3: BLEU evaluation of hypothesis selection

MT engine	BTEC^O	BTEC^V
SEL	**0.5409** (+ 0.7%)	**0.5470** (+ 3.6%)
ORACLE	0.6385 (+10.4%)	0.6502 (+13.9%)

MT engine	BTEC$^{O\cup V}$
SEL	**0.4648** (–7.0%)
ORACLE	0.6969 (+16.3%)

The results show that the selection method is effective for both variant corpora whereby a larger gain is achieved for BTEC^V. However, the ORACLE results indicate that the method fails to tap the full potential of the element MT engines.

In addition, we trained the statistical models of the hypothesis selection method on the corpus obtained

[2]The *word error rate* (WER) is an objective evaluation measures that, in contrast to BLEU, can be applied on sentence-level. It penalizes edit operations for the translation output against reference translations.

by merging all variant corpora (BTEC$^{O \cup V}$). Despite the larger amount of training data, the BLEU score decreases drastically which shows that an increase in training data not necessarily leads to improved translation quality. Moreover, the ORACLE selection applied to all translation hypotheses based on the BTEC^{O} as well as the BTEC^{V} corpus indicates that both variants can contribute significantly in order to improve the overall system performance.

4.3 Effects of Variant Selection

The effects of combining selected variant hypotheses by testing whether significant differences in statistical scores were obtained are summarized in Table 4. The variant selection method is applied to the translation outputs of each element MT engine ($\text{MT}_j^O \parallel \text{MT}_j^V$) as well as the selected translation hypotheses ($\text{MT}_{SEL}^O \parallel \text{MT}_{SEL}^V$). The gain of the proposed variant selection method relative the best element MT output based on a single variant corpus is given in parentheses.

Table 4: BLEU evaluation of variant selection

MT engine		BTEC^O $\parallel$ BTEC^V	
SMT	$\text{MT}_1^O \parallel \text{MT}_1^V$	0.5010	(+ 3.8%)
	$\text{MT}_2^O \parallel \text{MT}_2^V$	0.4847	(+ 2.5%)
	$\text{MT}_3^O \parallel \text{MT}_3^V$	0.5594	(+ 2.5%)
	$\text{MT}_4^O \parallel \text{MT}_4^V$	0.4733	(+ 2.7%)
EBMT	$\text{MT}_5^O \parallel \text{MT}_5^V$	0.3863	(+ 4.2%)
	$\text{MT}_6^O \parallel \text{MT}_6^V$	0.4338	(+ 2.4%)
	$\text{MT}_7^O \parallel \text{MT}_7^V$	0.0181	(+10.7%)
	$\mathbf{MT}_{SEL}^O \parallel \mathbf{MT}_{SEL}^V$	**0.5765**	(+ 4.2%)

The results show that the variant selection method is effective for all element MT engines. The highest BLEU score is achieved for $\mathbf{MT}_{SEL}^O \parallel \mathbf{MT}_{SEL}^V$ gaining 4.2% in BLEU score. Moreover, the proposed method outperforms the hypothesis selection method based on the merged corpus BTEC$^{O \cup V}$ by 11.2% in BLEU score.

A comparison of the proposed method with the best performing system (C-STAR data track, BLEU=0.5279) of the IWSLT 2005 workshop showed that our system outperforms the top-ranked system gaining 4.8% in BLEU score.

5 Conclusion

This paper proposed the usage of variant corpora to improve the translation quality of a multi-engine-based approach to machine translation. The element MT engines were used to translate the same input whereby the best translation was selected according to statistical models. A test on the significance of differences between statistical scores judging the translation quality of a given hypothesis was exploited to identify the model that fits the input sentence best and the respective translation hypothesis was selected as the translation output.

The proposed method was evaluated on the CE translation task of the IWSLT 2005 workshop. The results showed that the proposed method achieving a BLEU score of 0.5765 outperformed not only all element MT engines (gaining 3.6% in BLEU score), but also a selection method using a larger corpus obtained from merging all variant corpora (gaining 11.2% in BLEU score) due to less ambiguity in the utilized models. In addition, the proposed method also outperformed the best MT system (C-STAR data track) of the IWSLT 2005 workshop gaining 4.8% in BLEU score.

Further investigations should analyze the characteristics of the variant corpora in more detail and focus on the automatic identification of specific linguistic phenomena that could be helpful to measure how good an input sentence is covered by a specific model. This would allow us to select the most adequate variant beforehand, thus reducing computational costs and improving the system performance. This would also enable us to cluster very large corpora according to specific linguistic phenomena, thus breaking down the full training corpus to consistent subsets that are easier to manage and that could produce better results.

References

K. Papineni et al. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318.

Y. Akiba et al. 2002. Using language and translation models to select the best among outputs from multiple MT systems. In *Proc. of COLING*, pages 8–14.

G. Kikui et al. 2003. Creating corpora for speech-to-speech translation. In *Proc. of EUROSPEECH03*, pages 381–384.

M. Paul et al. 2005. Nobody is Perfect: ATR's Hybrid Approach to Spoken Language Translation. In *Proc. of the IWSLT*, pages 55–62.