Boosting Statistical Machine Translation by Lemmatization and Linear Interpolation

Ruiqiang Zhang[1,2] and **Eiichiro Sumita**[1,2]
[1]National Institute of Information and Communications Technology
[2]ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Seiika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{ruiqiang.zhang,eiichiro.sumita}@atr.jp

Abstract

Data sparseness is one of the factors that degrade statistical machine translation (SMT). Existing work has shown that using morpho-syntactic information is an effective solution to data sparseness. However, fewer efforts have been made for Chinese-to-English SMT with using English morpho-syntactic analysis. We found that while English is a language with less inflection, using English lemmas in training can significantly improve the quality of word alignment that leads to yield better translation performance. We carried out comprehensive experiments on multiple training data of varied sizes to prove this. We also proposed a new effective linear interpolation method to integrate multiple homologous features of translation models.

1 Introduction

Raw parallel data need to be preprocessed in the modern phrase-based SMT before they are aligned by alignment algorithms, one of which is the well-known tool, GIZA++ (Och and Ney, 2003), for training IBM models (1-4). Morphological analysis (MA) is used in data preprocessing, by which the surface words of the raw data are converted into a new format. This new format can be lemmas, stems, parts-of-speech and morphemes or mixes of these. One benefit of using MA is to ease data sparseness that can reduce the translation quality significantly, especially for tasks with small amounts of training data.

Some published work has shown that applying morphological analysis improved the quality of SMT (Lee, 2004; Goldwater and McClosky, 2005). We found that all this earlier work involved experiments conducted on translations from highly inflected languages, such as Czech, Arabic, and Spanish, to English. These researchers also provided detailed descriptions of the effects of foreign language morpho-syntactic analysis but presented no specific results to show the effect of English morphological analysis. To the best of our knowledge, there have been no papers related to English morphological analysis for Chinese-to-English (CE) translations even though the CE translation has been the main track for many evaluation campaigns including NIST MT, IWSLT and TC-STAR, where only simple tokenization or lower-case capitalization has been applied to English preprocessing. One possible reason why English morphological analysis has been neglected may be that English is less inflected to the extent that MA may not be effective. However, we found this assumption should not be taken-for-granted.

We studied what effect English lemmatization had on CE translation. Lemmatization is shallow morphological analysis, which uses a lexical entry to replace inflected words. For example, the three words, *doing*, *did* and *done*, are replaced by one word, *do*. They are all mapped to the same Chinese translations. As a result, it eases the problem with sparse data, and retains word meanings unchanged. It is not impossible to improve word alignment by using English lemmatization.

We determined what effect lemmatization had in experiments using data from the BTEC (Paul, 2006) CSTAR track. We collected a relatively large corpus of more than 678,000 sentences. We conducted comprehensive evaluations and used multiple trans-

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 181–184,
Prague, June 2007. ©2007 Association for Computational Linguistics

lation metrics to evaluate the results. We found that our approach of using lemmatization improved both the word alignment and the quality of SMT with a small amounts of training data, and, while much work indicates that MA is useless in training large amounts of data (Lee, 2004), our intensive experiments proved that the chance to get a better MT quality using lemmatization is higher than that without it for large amounts of training data.

On the basis of successful use of lemmatization translation, we propose a new linear interpolation method by which we integrate the homologous features of translation models of the lemmatization and non-lemmatization system. We found the integrated model improved all the components' performance in the translation.

2 Moses training for system with lemmatization and without

We used Moses to carry out the expriments. Moses is the state of the art decoder for SMT. It is an extension of Pharaoh (Koehn et al., 2003), and supports factor training and decoding. Our idea can be easily implemented by Moses. We feed Moses English words with two factors: surface word and lemma. The only difference in training with lemmatization from that without is the alignment factor. The former uses Chinese surface words and English lemmas as the alignment factor, but the latter uses Chinese surface words and English surface words. Therefore, the lemmatized English is only used in word alignment. All the other options of Moses are same for both the lemmatization translation and non-lemmatization translation.

We use the tool created by (Minnen et al., 2001) to complete the morphological analysis of English. We had to make an English part-of-speech (POS) tagger that is compatible with the CLAWS-5 tagset to use this tool. We use our in-house tagset and English tagged corpus to train a statistical POS tagger by using the maximum entropy principle. Our tagset contains over 200 POS tags, most of which are consistent to the CLAWS-5. The tagger achieved 93.7% accuracy for our test set.

We use the default features defined by Pharaoh in the phrase-based log-linear models i.e., a target language model, five translation models, and one distance-based distortion model. The weighting parameters of these features were optimized in terms of BLEU by the approach of minimum error rate training (Och, 2003).

The data for training and test are from the IWSLT06 CSTAR track that uses the Basic Travel Expression Corpus (BTEC). The BTEC corpus are relatively larger corpus for travel domain. We use 678,748 Chinese/English parallel sentences as the training data in the experiments. The number of words are about 3.9M and 4.4M for Chinese and English respectively. The number of unique words for English is 28,709 before lemmatization and 24,635 after lemmatization. A 15%-20% reduction in vocabulary is obtained by the lemmatization. The test data are the one used in IWSLT06 evaluation. It contains 500 Chinese sentences. The test data of IWSLT05 are the development data for tuning the weighting parameters. Multiple references are used for computing the automatic metrics.

3 Experiments

3.1 Regular test

The purpose of the regular tests is to find what effect lemmatization has as the amount of training data increases. We used the data from the IWSLT06 CSTAR track. We started with 50,000 (50 K) of data, and gradually added more training data from a 678 K corpus to this. We applied the methods in Section 2 to train the non-lemmatized translation and lemmatized translation systems. The results are listed in Table 1. We use the alignment error rate (AER) to measure the alignment performance, and the two popular automatic metric, BLEU[1] and METEOR[2] to evaluate the translations. To measure the word alignment, we manually aligned 100 parallel sentences from the BTEC as the reference file. We use the "sure" links and the "possible" links to denote the alignments. As shown in Table 1, we found our approach improved word alignment uniformly from small amounts to large amounts of training data. The maximal AER reduction is up to 27.4% for the 600K. However, we found some mixed translation results in terms of BLEU. The lemmatized

[1] http://domino.watson.ibm.com/library/CyberDig.nsf (keyword=RC22176)

[2] http://www.cs.cmu.edu/~alavie/METEOR

Table 1: Translation results as increasing amount of training data in IWSLT06 CSTAR track

System		AER	BLEU	METEOR
50K	nonlem	0.217	0.158	0.427
	lemma	0.199	0.167	0.431
100K	nonlem	0.178	0.182	0.457
	lemma	0.177	0.188	0.463
300K	nonlem	0.150	0.223	0.501
	lemma	0.132	0.217	0.505
400K	nonlem	0.136	0.231	0.509
	lemma	0.102	0.224	0.507
500K	nonlem	0.119	0.235	0.519
	lemma	0.104	0.241	0.522
600K	nonlem	0.095	0.238	0.535
	lemma	0.069	0.248	0.536

Table 2: Statistical significance test in terms of BLEU: sys1=non-lemma, sys2=lemma

Data size	Diff(sys1-sys2)
50K	-0.092 [-0.0176,-0.0012]
100K	-0.006 [-0.0155,0.0039]
300K	0.0057 [-0.0046,0.0161]
400K	0.0074 [-0.0023,0.0174]
500K	-0.0054 [-0.0139,0.0035]
600K	-0.0103 [-0.0201,-0.0006]

translations did not outperform the non-lemmatized ones uniformly. They did for small amounts of data, i.e., 50 K and 100 K, and for large amounts, 500 K and 600 K. However, they failed for 300 K and 400 K.

The translations were under the statistical significance test by using the *bootStrap* scripts[3]. The results giving the medians and confidence intervals are shown in Table 2, where the numbers indicate the median, the lower and higher boundary at 95% confidence interval. we found the *lemma* systems were confidently better than the *nonlem* systems for the 50K and 600K, but didn't for other data sizes.

This experiments proved that our proposed approach improved the qualities of word alignments that lead to the translation improvement for the 50K, 100K, 500K and 600K. In particular, our results revealed large amounts of data of 500 K and 600 K was improved by the lemmatization while it has been found impossible in most published results. However, data of 300 K and 400 K worsen translations achieved by the lemmatization[4]. In what follows, we discuss a method of random sampling of creating multiple corpora of varied sizes to see robustness of our approach and re-investigate the results of the 300K and 400K.

Table 3: Competitive scores (BLEU) for non-lemmatization and lemmatization using randomly extracted corpora

System	100K	300K	400K	600K	total
lemma	10/11	5.5/11	6.5/11	5/7	27/40
nonlem	1/11	5.5/11	4.5/11	2/7	13/40

3.2 Random sampling test

In this section, we use a method of random extraction to generate new multiple training data for each corpus of one definite size. The new data are extracted from the whole corpus of 678 K randomly. We generate ten new corpora for 100 K, 300 K, and 400 K data and six new corpora for the 678 K data. Thus, we create eleven and seven corpora of varied sizes if the corpora in the last experiments are counted. We use the same method as in Section 2 for each generated corpus to construct systems to compare non-lemmatization and lemmatization. The systems are evaluated again using the same test data. The results are listed in Table 3 and Figure 1. Table 3 shows the "scoreboard" of non-lemmatized and lemmatized results in terms of BLEU. If its score for the *lemma* system is higher than that for the *nonlem* system, the former earns one point; if equal, each earns 0.5; otherwise, the *nonlem* earns one point. As we can see from the table, the results for the *lemma* system are better than those for the *nonlem* system for the 100K in 10 of the total 11 corpora. Of the total 40 random corpora, the *lemma* systems outperform the *nonlem* systems in 27 times.

By analyzing the results from Tables 1 and 3, we can arrive at some conclusions. The *lemma* systems outperform the *nonlem* for training corpora less than

[3] http://projectile.is.cs.cmu.edu/research/public/tools/bootStrap /tutorial.htm

[4] while the results was not confident by statistical significance test, the medians of 300K and 400K were lowered by the lemmatization

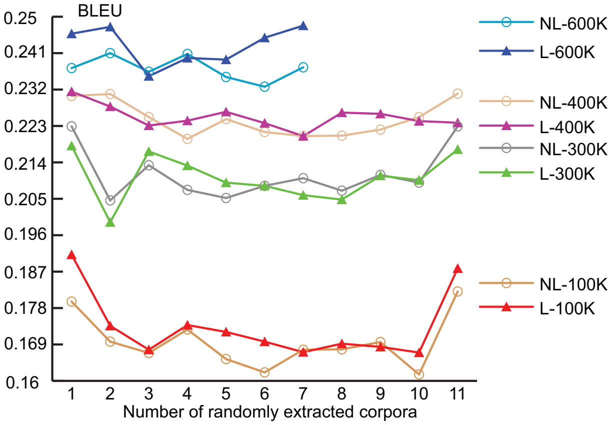

Figure 1: Bleu scores for randomly extracted corpora

100 K. The BLEU score favors the *lemma* system overwhelmingly for this size. When the amount of training data is increased up to 600 K, the *lemma* still beat the *nonlem* system in most tests while the number of success by the *nonlem* system increases. This random test, as a complement to the last experiment, reveals that the *lemma* either performs the same or better than the *nonlem* system for training data of any size. Therefore, the *lemma* system is slightly better than the *nonlem* in general.

Figure 1 illustrates the BLEU scores for the "lemma(L)" and "nonlem(NL)" systems for randomly extracted corpora. A higher number of points is obtained by the *lemma* system than the *nonlem* for each corpus.

4 Effect of linear interpolation of features

We generated translation models for lemmatization translation and non-lemmatization translation. We found some features of the translation models could be added linearly. For example, phrase translation model $p(e|f)$ can be calculated as,

$$p(e|f) = \alpha_1 p_l(e|f) + \alpha_2 p_{nl}(e|f)$$

where $p_l(e|f)$ and $p_{nl}(e|f)$ is the phrase translation models corresponding to the lemmatization system and non-lemma system. $\alpha_1 + \alpha_2 = 1$. αs can be obtained by maximizing likelihood or BLEU scores of a development data. But we used the same values for all the α. $p(e|f)$ is the phrase translation model after linear interpolation. Besides the phrase translation model, we used this approach to integrate the three other features: phrase inverse probability, lexical probability, and lexical inverse probability. We tested this integration using the open track of IWSLT 2006, a small task track. The BLEU scores are shown in Table 4. An improvement over both of the systems were observed.

Table 4: Effect of linear interpolation

	lemma	nonlemma	interpolation
open track	0.1938	0.1993	0.2054

5 Conclusions

We proposed a new approach of using lemmatization and linear interpolation of homologous features in SMT. The principal idea is to use lemmatized English for the word alignment. Our approach was proved effective for the BTEC Chinese to English translation. It is significant in particular that we have target language, English, as the lemmatized object because it is less usual in SMT. Nevertheless, we found our approach significantly improved word alignment and qualities of translations.

References

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of HLT/EMNLP*, pages 676–683, Vancouver, British Columbia, Canada, October.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL 2003: Main Proceedings*, pages 127–133.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *HLT-NAACL 2004: Short Papers*, pages 57–60, Boston, Massachusetts, USA.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL 2003*, pages 160–167.

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proc. of the IWSLT*, pages 1–15, Kyoto, Japan.