On the Use of Lemmatization for Statistical Machine Translation

Ruiqiang Zhang[1,2] and **Hirofumi Yamamoto** 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

We propose using English lemmatization to improve statistical machine translation (SMT) from Chinese to English. We implemented an English part-of-speech tagger for CLAW-5 tag set, and use an English morphological analyzer to lemmatize English. Our approach is proved very effective for SMT with small amount of training data. When we applied our approach for large amount of training data, we found that our approach improved SMT in most experiments while the approach was not effective in a few cases. We carried out comprehensive experiments and used the state-of-the-art evaluation metrics (BLEU, METEOR and TER) to derive the conclusions.

1 Introduction

In modern phrase-based SMT, the raw bilingual corpus need to be preprocessed before the parallel data are aligned by some alignment algorithms, for example, using the well-known tool (Och and Ney, 2003), GIZA++, training the IBM models (1-4). Morphological analysis (MA) is used in the data preprocessing, by which the surface word format of raw data is 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. Data sparseness is one of the factors to influence the SMT performance. Especially for the task with small training data, it can reduce the translation quality significantly.

It has been shown in some published work that applying morphological analysis improved SMT quality (Lee, 2004; Sadat and Habash, 2006; Goldwater and McClosky, 2005; Gupta and Federico, 2006). However, few work has been reported about English morphological analysis for Chinese to English (CE) translation even though the CE translation is a main task for many evaluations including NIST MT, IWSLT and TCSTAR, where only a simple tokenization is applied for English preprocessing. One possible reason why English morphological analysis was ignored may be that English language is such a less inflected language that MA may not be effective. However, our work shows this assumption cannot be taken for granted.

We studied the effect of English lemmatization for CE translation. The lemmatization is a shallow morphological analysis, which uses the lexicon entry to replace the inflected format of words. For example, the three words, doing, did and done, are replaced by one word,do. The lemmatization normalizes many inflected words into one lexical entry. As a result, the lemmatization reduced data sparseness.

We determined what effect lemmatization had in experiments using data from the BTEC (Paul, 2006) OPEN track and NIST MT05 task. We conducted comprehensive evaluations and used multiple trans-

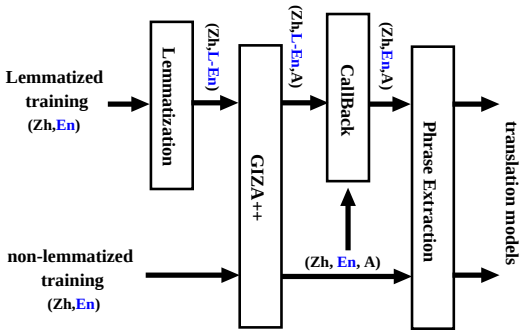

Figure 1: Comparison of the two systems

lation metrics to evaluate the results. We found that our approach of using lemmatization improved the quality of SMT with a small amounts and large amounts of training data even though much work indicates that MA is useless in training large amounts of data (Gupta and Federico, 2006; Lee, 2004).

2 A system comparison for training of lemmatization and non-lemmatization

Figure. 1 illustrates differences of the non-lemmatization translation and lemmatization translation. The top lines show the training process of non-lemmmatization translation while the bottom lines indicate the lemmatization translation. Two additional steps, *Lemmatization* and *CallBack*, were used in the lemmatizaton translation. The *Lemmatization* functions as to lemmatizing English before the parallel corpus is sent to the GIZA++ training for word alignment. The *CallBack* part replaced the lemmatized English with the original non-lemmatized English after the word alignment was finished. Hence, lemmatization was used only for word alignment.

We used the tool by (Minnen et al., 2001) to complete the morphological analysis for English. We have to make an English part-of-speech (POS) tagger compatible with the CLAWS-5 tagset in order to use the tool. We implemented the tagger by maximum entropy principle. Our tagset contains over 200 POS tags, most of which is consistent to the CLAWS-5. We have manually labeled one million English corpus with the POS tagset. The POS tagger achieved 93.7% accuracy for our test set.

We used the Pharaoh decoder that is a beam search decoding process for maximizing a feature based log-linear models. We used the default features defined by the Pharaoh: target language model, five translation models and one internal distance based distortion model.

Table 1: Statistics of data in use (sentence number and vocabulary size)

		OPEN	MT05
Train	#sent.	39,953	2,399,753
	#lemma	7,726	208,021
	#nonlem	9,207	253,933
Test	#sent.	500	1082

3 Data

We used the data from IWSLT06 OPEN and NIST MT05 in our experiments. The data statistics were shown in Table 3. IWSLT06 OPEN track used a small amount of training data (see Table 3). NIST MT05 consists of a very huge amount of training data. we used the correct recognition reference data in the OPEN track. In the Table 3, *lemma* stands for the corpus by lemmatization; *nonlem*, the corpus without using the lemmatization. The table shows the size of vocabulary in the *lemma* and *nonlem* system. A 15%-20% vocabulary reduction was observed by the lemmatization.

4 Experiments

In this section we used the data from IWSLT06 OPEN track and NIST MT05 large track. The OPEN track uses a small training data and MT05 uses a very large data as shown in Table 3. The results are shown in Table 4. The lemmatization gave significant improvement to the OPEN track with a small amount of data, but slightly improved the NIST MT05 track with large amount of training data.

5 Discussion and conclusions

We studied the use of lemmatization for Chinese to English SMT. As we stated in the introduction, the work of this paper has not been tried while there exists some published work on morphological analysis for SMT in other languages except the CE. Our re-

Table 2: Results in IWSLT06 OPEN track and NIST MT05 large track

		BLEU	NIST	METEOR	TER
OPEN	lemma	0.196	6.15	0.474	63.34
	nonlem	0.191	6.15	0.468	64.13
MT05	lemma	0.2457	8.01	0.537	67.64
	nonlem	0.2434	7.98	0.535	68.65

search results prove the positive effects of using the lemmatization in both the small data and large data.

Lemmatization is the simplest MA approach. A deep MA approach is to integrate the lemma with all the available morphemes such as parts-of-speech, stems and word senses. A deep combination in MA is the focus for our future work.

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.

Deepa Gupta and Marcello Federico. 2006. Exploiting word transformation in statistical machine translation from spanish to english. In *Proceedings of the 11th EAMT*, Oslo, Norway, June.

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.

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

Fatiha Sadat and Nizar Habash. 2006. Combination of arabic preprocessing schemes for statistical machine translation. In *Proceedings of the COLING/ACL*, pages 1–8, Sydney, Australia, July.