Effects of Integrating Multiple Bilingually-Trained Segmentation Schemes for Japanese-English SMT

Michael Paul and Andrew Finch and Eiichiro Sumita
MASTAR Project
National Institute of Information and Communications Technology
Hikaridai 3-5, Keihanna Science City
619-0289 Kyoto, Japan
michael.paul@nict.go.jp

Abstract

This paper proposes a method to integrate multiple segmentation schemes into a single statistical machine translation (SMT) system by characterizing the source language side and merging identical translation pairs of differently segmented SMT models. Experimental results translating Japanese into English revealed that the proposed method of integrating multiple segmentation schemes outperforms SMT models trained on any of the utilized word segmentations and performs comparably to available state-of-the-art monolingually built segmentation tools.

1 Introduction

The task of *word segmentation*, i.e., identifying word boundaries in continuous text, is one of the fundamental preprocessing steps of data-driven NLP applications like *Natural Language Understanding* or *Machine Translation*. In contrast to Indo-European languages like *English*, many Asian languages like *Japanese* do not use a whitespace character to separate meaningful word units. The problems of word segmentation are:

(1) *ambiguity*, e.g., for Japanese, a single character can be a word component in one context, but a word by itself in another context.

(2) *unknown words*, i.e., existing words can be combined into new words not existing in the training data, e.g., the combination of "白" ("white") and "鳥" ("bird") should be translated as "*swan*" and not as "*white bird*".

Purely dictionary-based approaches like (Cheng et al., 1999) addressed these problems by maximum matching heuristics, but their accuracy depends largely on the coverage in the utilized dictionary. Recent research on word segmentation focuses on approaches based on probabilistic methods (Brent, 1999; Goldwater et al., 2006). For machine translation applications, improvements have been reported for approaches taking into account not only monolingual, but also bilingual information to derive a word segmentation suitable for SMT. For example, (Xu et al., 2008) proposes a Bayesian Semi-Supervised approach for Chinese word segmentation that builds on (Goldwater et al., 2006).

Recent research on SMT is also focusing on the usage of multiple word segmentation schemes on the source language to improve translation quality. For example, (Zhang et al., 2008) combines dictionary-based and CRF-based approaches for Chinese word segmentation in order to avoid *out-of-vocabulary* (OOV) words. Moreover, (Nakov et al., 2009) utilizes SMT engines trained on different word segmentation schemes and combines the translation outputs using system combination techniques as a post-process to SMT decoding.

The method proposed in this papers integrates multiple word segmentation scheme information directly into the SMT decoding process. For each of the word segmentation schemes, a standard SMT engine is built and the statistical translation models are merged by characterizing the source side of each translation model, summing up the probabilities of identical phrase translation pairs, and rescoring the merged translation model (cf. Section 2).

The proposed method is applied to the translation of *Japanese* into *English*. The utilized language resources and the outline of the experiments are summarized in Section 3. The experimental results revealed that the proposed method outperforms not only a baseline system that translates characterized source language sentences but also all SMT models trained on word segmentations automati-

cally learned using a parallel text corpus. In addition, the proposed method achieves translation results comparable to SMT models trained on bitext segmented with linguistic tools.

2 Integration of Multiple Word Segmentation Schemes

The proposed method is language-independent and can handle any type of word segmentations scheme. For the experiments in Section 3, we are using different word segmentation schemes that are learned automatically using a parallel corpus by (1) aligning source language sentences character-wise to word units separated by a whitespace in the target language and (2) applying an iterative bootstrap algorithm to learn larger source language units which optimizes translation quality (Paul et al., 2009).

The integration of multiple word segmentation schemes into a single SMT engine is carried out by merging the statistical models of SMT engines trained on the characterized and iteratively learned word segmentation schemes, i.e., the model probabilities of identical source/target phrase translation pairs are summed up. Concerning the target language part, exact matches can be directly obtained, because the target language phrases of all the iteration models are segmented using the same word segmentation scheme. However, the segmentation of source language phrases can differ between the iterative models. In order to allow a full exact match, the source language side of all translation pairs of each model is characterized prior to the merging step. After merging, the obtained statistical translation models have to be rescored to get a normalized score representing the translation probability of the merged source/target phrase translation pairs.

The rescored translation model covers all translation pairs that were learned by any of the iterative models. Therefore, the selection of longer translation units during decoding can reduce the complexity of the translation task, if applicable. On the other hand, overfitting problems of single-iteration models can be avoided because multiple smaller source language translation units can be exploited to cover the given source language input parts and to generate translation hypotheses based on the concatenation of associated target phrase expressions. Moreover, the merging process increases the translation probabilities of those source/target translation parts that cover the same surface string, but differ only in the segmentation of the source language phrase. Therefore, the more often such a translation pair is learned by different iterative models, the more often the respective target language expression will be exploited by the SMT decoder.

The translation of unseen data using the merged translation models is carried out by (1) characterizing the source language input text and (2) applying the SMT decoding in a standard way.

3 Experiments

The effects of using different word segmentations and integrating them into an SMT engine are investigated using the multilingual *Basic Travel Expressions Corpus* (BTEC), which is a collection of sentences that bilingual travel experts consider useful for people going to or coming from other countries (Kikui et al., 2006). Table 1 summarizes the characteristics of the BTEC corpus used for the training of the SMT models (*train*), the tuning of model weights (*dev*), and the evaluation of translation quality (*eval*). Besides the number of sentences (*sen*) and the vocabulary (*voc*), the sentence length (*len*) is also given.

The given statistics are obtained using *ChaSen*[1], a linguistic Japanese segmentation tool, and a simple tokenization script separating punctuation marks in the English data sets.

BTEC		train set	dev set	eval set
# of sen		160,000	1,000	1,000
en	voc	15,390	1,262	1,292
	len	7.5	7.1	7.2
ja	voc	17,168	1,407	1,408
	len	8.5	8.2	8.2

Table 1: Language Resources

For the training of the SMT models, standard word alignment (Och and Ney, 2003) and language modeling (Stolcke, 2002) tools were used. Minimum error rate training (MERT) was used to tune the decoder's parameters, and performed on the *dev* set using the technique proposed in (Och and Ney, 2003). For the translation, an in-house multi-stack phrase-based decoder comparable to the open-source toolkit MOSES was used. For the evaluation

[1] http://chasen.naist.jp/hiki/ChaSen

of translation quality, we applied standard automatic evaluation metrics, i.e., BLEU (Papineni, 2002) and METEOR (Banerjee and Lavie, 2005). For the experimental results in this paper, the given scores are listed as percentage figures.

In addition, human assessment of translation quality was carried out using the *Ranking* metrics. For the *Ranking* evaluation, a human grader was asked to *"rank each whole sentence translation from Best to Worst relative to the other choices (ties are allowed)"* (Callison-Burch et al., 2007). The *Ranking* scores were obtained as the average number of times that a system was judged better than any other system and the normalized ranks (*NormRank*) were calculated on a per-judge basis for each translation task using the method of (Blatz et al., 2003).

The automatic evaluation scores of the SMT engines trained on the differently segmented source language resources are given in Table 2, where:

character	refers to the baseline system of using character segmented source text for the translation.
single-best	is the SMT engine that is trained on the corpus segmented by the best-performing iteration of the bootstrap approach.
proposed	is the SMT engine whose translation models integrate multiple word segmentation schemes.
linguistic	uses the linguistically motivated word segmentation tool *ChaSen*.

The automatic evaluation results (BLEU, METEOR) show, that the proposed method outperforms the *character* (*single-best*) system for each of the involved languages for both evaluation metrics achieving gains of 4.5 (2.6) BLEU points and 4.2 (1.3) METEOR points, respectively. Comparing the proposed method towards the linguistically motivated segmenter, the results show that slightly lower automatic evaluation scores were achieved for the integrated word segmentations for Japanese, although the results of the proposed method are quite close.

The preliminary subjective evaluation results were carried out on a randomly selected subset of 600 input sentences by a paid evaluation expert who is a native speaker of English. The *RankNorm* results confirm mainly the findings of the automatic evaluation. However, the translation outputs of the proposed method were judged better than those of the linguistically segmented SMT model.

source language	word segmentation			
	character	single-best	proposed	linguistic
BLEU	40.14	42.13	44.70	**44.99**
METEOR	60.05	63.04	64.36	**64.88**
NormRank	2.76	2.85	**3.18**	3.12

Table 2: Japanese-English Translation Quality

Table 3 illustrates some translation examples using different segmentation schemes for the Japanese-English translation task. The SMT engines that output the best translations are marked with an asterisk. In the first example, the concatenation of "もう真夜中" (*already midnight*) by the *single-best* segmentation scheme leads to an OOV word, thus only a partial translation can be achieved. However, the problem can be resolved using the proposed method. The second example is best translated using the *single-best* word segmentation that correctly handles the sentence coordination. The proposed method generates an additional conjunction, but the coordinated sentence parts are translated correctly. The baseline system omits the sentence coordination information resulting in an unacceptable translation. The third examples illustrates that longer tokens reduce the translation complexity and thus can be translated better than the other segmentation that cause more ambiguities.

4 Conclusions

This paper proposed a new language-independent method to intergrate multiple word segmentation schemes of languages that do not use whitespace characters to separate meaningful word units. The proposed method can handle any type of word segmentations scheme. The effectiveness of the proposed method was investigated for the translation of *Japanese* into *English* for the domain of travel conversations. The automatic evaluation of the translation results showed consistent improvements compared to a baseline system that translates characterized input sentences and the best performing SMT engine of the iterative learning procedure, respectively. In addition, the proposed method achieved translation results similar to SMT models trained on bitext segmented with linguistically-motivated tools,

Table 3: Sample Translations

linguistic	*seg:* ええ/。/えーと/、/もう/真夜中/です/ね 。 *trans:* Yes. Let's see. It's midnight.
character*	*seg:* え/え/。/え/ー/と/、/も/う/真/夜/中/で/す/ね/。 *trans:* Yes. Well, it's already midnight.
single-best	*seg:* ええ。/えーと/、/もう真夜中/で/すね。 *trans:* Yes. Let 's see.
proposed*	*seg:* え/え/。/え/ー/と/、/も/う/真/夜/中/で/す/ね/。 *trans:* Yes. Well, it's already midnight.
linguistic	*seg:* ジーンズ/が/欲しい/の/です/が/、/ いい/店/を/教え/て/ください/。 *trans:* I'd like a pair of jeans. Could you recommend a good shop?
character	*seg:* ジ/ー/ン/ズ/が/欲/し/い/の/で/す/が/、/ い/い/店/を/教/え/て/く/だ/さ/い/。 *trans:* Could you recommend a good 'd like a pair of jeans.
single-best*	*seg:* ジーンズ/が欲し/い/のですが、/ いい/店/を教え/てください/。 *trans:* I'd like some jeans. Could you recommend a good shop?
proposed	*seg:* ジ/ー/ン/ズ/が/欲/し/い/の/で/す/が/、/ い/い/店/を/教/え/て/く/だ/さ/い/。 *trans:* I 'd like a pair of jeans and could you recommend a good shop?
linguistic	*seg:* 今日/の/午後/まで/に/でき/ます/か/。 *trans:* Will it be ready by this afternoon?
character	*seg:* 今/日/の/午/後/ま/で/に/で/き/ま/す/か/。 *trans:* It'll be ready by this afternoon?
single-best	*seg:* 今日の/午後まで/に/できま/すか。 *trans:* Will it be ready by this afternoon?
proposed*	*seg:* 今/日/の/午/後/ま/で/に/で/き/ま/す/か/。 *trans:* Can you have these ready by this afternoon?

although no external information but only the given bitext was used to train the segmentation models.

For Japanese, which is written using three different scripts (*kanji*, *hiragana*, *katakana*), additional features in the script type of a given token might also help to improve the translation quality of SMT systems trained on automatically learned word segmentation schemes, thus improving the performance of the proposed integration method further.

Acknowledgment

This work is partly supported by the Grant-in-Aid for Scientific Research (C) No. 19500137.

References

S. Banerjee and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation. In *Proc. of the ACL*, pages 65–72, Ann Arbor, US.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2003. Confidence estimation for statistical machine translation. In *Final Report of the JHU Summer Workshop*.

M. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on SMT*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

K. Cheng, H. Young, and K. Wong. 1999. A study on word-based and integrat-bit Chinese text compression algorithms. *American Society of Information Science*, 50(3):218–228.

S. Goldwater, T. Griffith, and M. Johnson. 2006. Contextual Dependencies in Unsupervised Word Segmentation. In *Proc. of the ACL*, pages 673–680, Sydney, Australia.

G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language*, 14(5):1674–1682.

P. Nakov, C. Liu, W. Lu, and H.T. Ng. 2009. The NUS SMT System for IWSLT 2009. In *Proc. of IWSLT*, pages 91–98, Tokyo, Japan.

F. Och and H. Ney. 2003. A Systematic Comparison of Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

K. Papineni. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th ACL*, pages 311–318, Philadelphia, US.

M. Paul, A. Finch, and E. Sumita. 2009. Language Independent Word Segmentation for Statistical Machine Translation. In *Proc. of the IUCS*, pages 36–40, Tokyo, Japan.

A. Stolcke. 2002. SRILM an extensible LM toolkit. In *Proc. of ICSLP*, pages 901–904, Denver, US.

J. Xu, J. Gao, K. Toutanova, and H. Ney. 2008. Bayesian Semi-Supervised Chinese Word Segmentation for SMT. In *Proc. of the COLING*, pages 1017–1024, Manchester, UK.

R. Zhang, K. Yasuda, and E. Sumita. 2008. Improved Statistical Machine Translation by Multiple Chinese Word Segmentation. In *Proc. of the Third Workshop on SMT*, pages 216–223, Columbus, Ohio, June. Association for Computational Linguistics.