# Chinese Word Segmentation Adaptation for Statistical Machine Translation

Click to edit Master subtitle style

Hailong Cao, Masao Utiyama and Eiichiro Sumita
Language Translation Group
NICT&ATR

# Introduction

- Chinese word segmentation (CWS) is a necessary step in Chinese-English statistical machine translation (SMT)
- Performance of CWS has an impact on the results of SMT.
- The common solution in Chinese-to-English translation has been to segment the Chinese text using an off-the-shelf CWS tool which is trained on manually segmented corpus.

# Main problems of using an off-the-shelf CWS tool

- Word granularity in the existing corpus is not necessarily suitable for SMT.
  - none of the existing corpora is specially developed for SMT
- When the CWS tool is used in a special domain which is different from its training corpus, the disambiguation ability of the CWS tool will drop and the performance of the SMT system will be influenced.

# Clues to solve the problems

- When we use a CWS tool to segment the Chinese side of Chinese-English parallel corpus which is used to train the SMT model, ***the English side is often be neglected.***
- Actually, there are many clues in the English side which can be used to determine an appropriate word granularity and resolve CWS ambiguity.

# Our solution: Adaptation

- We use two state-of-the-art CWS tools to preprocess the Chinese texts for our SMT system.
  - ICTCLAS (ICT, China)
  - hybrid model (NICT,ATR)
- Resolve CWS ambiguity by the information acquired by performing Chinese character to English word alignment by GIZA++ toolkits

# Adaptation algorithm for CWS

- For each sentence *C* in the Chinese side in the parallel corpus
- {
- W1 = segment *C* with ICTCLAS;
- W2 = segment *C* with the hybrid model;
- If ( W1 = = W2 )
- add W1 into the training set of the hybrid model;
- Else {
- A = character to word alignment of *C* and its English translation;
- W3= resolve (W1 ,W2 ,A);
- add W3 into the training set of the hybrid model;
- }
- }
- Retrain the hybrid model with the augmented data;

# An example

- There are two possible ways to segment the character sequence “马上来” in the Chinese sentence: “有人受伤了请马上来” (there has been a injury please come right away):
  - “马上 (right away) + 来 (come)”
  - “马 (horse) + 上来 (come up)”.
- All these four words are frequently used in Chinese text and it is very difficult for any CWS tool to make right decision without enough training data.

# An example (Cont.)

- The alignment result of the above sentence pair:
  - 有 -1 人 -2 受 -3 伤 -4 了 -5 请 -6 马 -7 上 -8 来 -9
  - NULL ({ }) there ({ 1 }) has ({ }) been ({ }) a ({ }) injury ({ 2 3 4 5 }) please ({ 6 }) come ({ 9 }) right ({ }) away ({ 7 8 })
- It is clear that “ 马 -7 ” and “ 上 -8 ” are aligned to the same English word “away”, while “ 来” is aligned to the word “come”. So we choose “ 马上 (right away) 来 (come)” as the right segmentation result.

# Experiments

- To evaluate the effect of our CWS adaptation algorithm, we apply it to the Chinese to English translation task of the IWSLT 2008.
- For comparison, we use three CWS tools.
  - ICTCLAS
  - hybrid model
  - Re-trained hybrid model

# Experimental setting

- Our SMT system is based on a fairly typical phrase-based model (Finch and Sumita, 2008).
- We use a 5-gram language model trained with modified Knesser-Ney smoothing.
- Minimum error rate training (MERT) with respect to BLEU score is used to tune the decoder's parameters

## hybrid model

| | BLEU | NIST | WER | METEOR | (BLEU+METEOR)/2 |
|---|---|---|---|---|---|
| Devset3 | 0.4749 | 8.5274 | 0.4280 | 0.7036 | 0.5893 |
| Devset5 | 0.1818 | 5.2429 | 0.7123 | 0.4430 | 0.3124 |
| Devset6 | 0.2551 | 5.3608 | 0.5826 | 0.5074 | 0.3813 |

## ICTCLAS

| | BLEU | NIST | WER | METEOR | (BLEU+METEOR)/2 |
|---|---|---|---|---|---|
| Devset3 | **0.4893*** | 8.3633 | **0.4072*** | 0.6985 | 0.5939 |
| Devset5 | 0.1826 | 4.7495 | 0.7042 | 0.4376 | 0.3101 |
| Devset6 | 0.2677 | 5.2655 | 0.5880 | 0.5067 | 0.3872 |

## Re-trained hybrid model

| | BLEU | NIST | WER | METEOR | (BLEU+METEOR)/2 |
|---|---|---|---|---|---|
| Devset3 | 0.4885 | **8.7183*** | 0.4273 | **0.7053*** | **0.5969*** |
| Devset5 | **0.1879*** | **5.2688*** | **0.6962*** | **0.4566*** | **0.3222*** |
| Devset6 | **0.2737*** | **5.5852*** | **0.5730*** | **0.5210*** | **0.3973*** |

# Related work

- Xu et al. (2005) ***integrate*** the segmentation process with the search for the best translation.
- Xu et al. (2006) propose an ***integration algorithm*** of English-Chinese word segmentation and alignment.
- Ma et al. (2007) introduce a method to pack words for word alignment.
- Chang et al. (2008) propose an algorithm to directly optimize segmentation granularity for translation quality.

# Conclusion

- A very simple and effective adaptation algorithm is proposed.
- Experimental results show that the our method can lead to better performance than two state-of-the-art CWS tools.

# Future work

- Now only two segmentation candidates are considered for each sentence In the future, we should extend our method to deal with n-best segmentation to get larger room for improvements.
- Now we simply combined all the Sighan corpora which adopt various specifications. So there should be inconsistent word granularity. We plan to acquire a uniform specification by making use of alignment information.

Any comment is welcome!

谢谢！