1. Abstract

In this study, the TÜBİTAK-UEKAE statistical machine translation system based on the open-source phrase-based statistical machine translation software, Moses, is presented. Additionally, phrase-table augmentation is applied to maximize source language coverage; lexical approximation is applied to replace out-of-vocabulary words with known words prior to decoding; and automatic punctuation insertion is improved. We describe the preprocessing and postprocessing steps and our training and decoding procedures.

2. Introduction

Among the six translation tasks in IWSLT 2008, we participated in the following:

- Arabic-to-English (BTEC Task)
- Chinese-to-English (BTEC Task)
- Chinese-to-Spanish (BTEC Task)
- Chinese-to-English-to-Spanish (Pivot Task)

Used Resources:

- Supplied training data
- Buckwaltor Arabic Morphological Analyzer (for BTEC AR-EN Task)

Lexical Approximation:

- To handle previously unseen words during decoding, the run-time lexical approximation method is used.
- An out-of-vocabulary word is replaced with the closest known word having the same feature.

This system obtained the best translation results in last year’s evaluation campaign (both AR-EN and JP-EN).

3. Training

Inclusion of Development Sets in Training
devsets1-3 were included in training (with references) in order to:

- Obtain better phrase alignments
- Increase the systems target phrase coverage

Sentence Splitting

Before translation model training, multi-sentence segments are split so as to prevent erroneous word alignments across sentence boundaries.

Orthographical Normalization

One of our goals from last year was to investigate the striking discrepancy between the performance of our system in correct recognition result (CRR) and ASR output conditions in the Arabic-to-English task.

- Also, /ı/ and /ı/ never occurred at the beginning of a word.

Phrase Table Augmentation

There may be some source-language words in the training corpus without a one-word entry in the phrase table. These words are treated as out-of-vocabulary in previously unseen contexts.

- Also, /ı/ and /ı/ never occurred at the beginning of a word.

Word pairs with lexical translation information is automatically restored.

4. Decoding

For decoding, Moses is used, which is a phrase-based beam-search decoder that uses a log-linear model with default scoring functions.

Run-time Lexical Approximation

The basic premise of lexical approximation is to replace a previously unseen word with a known word that has the same feature.

5. Results and Discussion

We have presented our Arabic-to-English, Chinese-to-English, Chinese-to-Spanish, and Chinese-to-English-to-Spanish statistical machine translation systems based on publicly-available software. We described our modifications to translation model generation, automatic punctuation insertion, and treatment of OOV words and presented our training and decoding procedures. Official evaluation results with correct recognition result and ASR output conditions were reported and discussed.