The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2008

Coşkun Mermer, Hamza Kaya, Ömer Farukhan Güneş, Mehmet Uğur Doğan

National Research Institute of Electronics and Cryptology (UEKAE), The Scientific and Technological Research Council of Turkey (TÜBİTAK), Gebze, Kocaeli 41470, Türkiye

(coskun,hamzaky,farukhan,mugur)@uekae.tubitak.gov.tr

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.

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

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
- Buckwalter 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.

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.

Task	Corpora	Sentence pairs	#segments
BTEC_AE	train DS1-3	44,164	**49,325**
BTEC_CE	train, DS1-3	44,164	**49,277**
BTEC_CS	train	19,972	**23,308**
PIVOT_CE	train	20,000	**22,563**
PIVOT_ES	train	19,972	**23,856**

Number of segments in the training corpora before and after automatic splitting

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.

Input condition	BLEU	Rank
Correct recognition result	49.23	1/11
ASR output	36.79	8/10

Official BLEU scores of the submitted AR-EN system in IWSLT 2007

In the supplied Arabic corpora;

- 8 Arabic characters / َ / ِ / ُ / ً / ٍ / ٌ / ْ / ّ / that were present in the training corpus were never used in the developments sets for the ASR output condition.

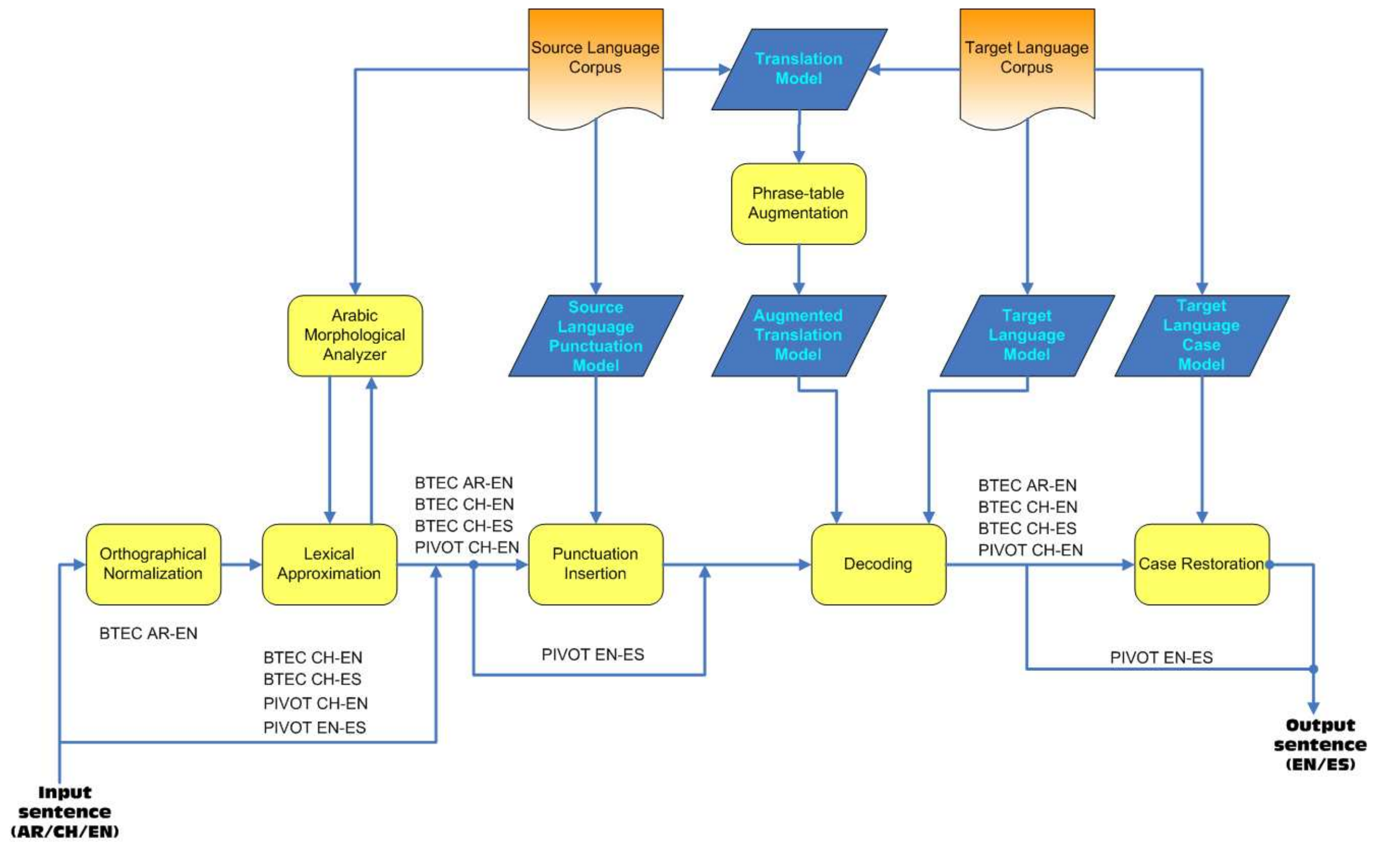

Figure 1: *IWSLT 2008 - TUBITAK UEKAE System.*

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

⟹ In order to match the ASR output, we orthographically normalized the training corpus by;

1. Removing all occurrences of the mentioned 8 characters.
2. Replacing all occurrences of / آ / and word-initial occurrences of / أ /and / إ / with / ا / (*alef*).

		devset4	devset5	devset6
ASR	Original orthography	23.14	19.96	37.67
	Normalized orthography	**23.95**	**20.29**	**41.32**
CRR	Original orthography	26.33	21.11	48.08
	Normalized orthography	**27.08**	**22.17**	**48.85**

Effect of orthographical normalization on ASR output and CRR translation BLEU scores in the BTEC AE task

We also tried this normalization for CRR translation. The table above shows the results for both ASR and CRR conditions. BLEU scores were improved in all development sets.

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.

⟹ Add such words as new phrase-pairs to the list of extracted phrases.

- The target phrases in these phrase-pairs are selected from GIZA++ word alignments.
- Word pairs with lexical translation probabilities above a relative threshold are selected.

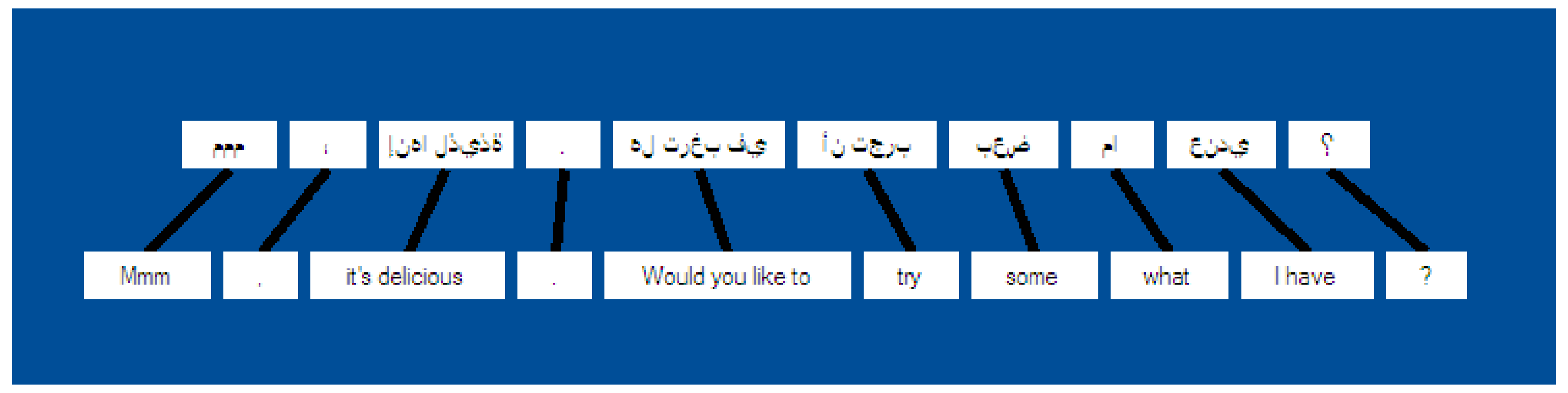

Figure 2: *Example Translation Output of the TÜBİTAK-UEKAE System.*

Corpus	BTEC			PIVOT	
	AE	CE	CS	CE	CS
$\|vcb\|$	17,720	8,757	8,412	9,186	7,074
$\|pt\|$	410K	395K	217K	216K	302K
$\|vcb_{miss}\|$	7,626	4,158	4,539	5,321	1,688
$\|\triangle pt\|$	20,610	13,190	16,619	21,122	3,754
$\|pt_{aug}\|$	**430K**	**408K**	**234K**	**237K**	**306K**

Phrase table augmentation. $|vcb|$: Source vocabulary size. $|pt|$: Default phrase table size. $|vcb_{miss}|$: Portion of source vocabulary without a one-word entry in the default phrase table. $|\triangle pt|$: New phrase-pairs added to the phrase table. $|pt_{aug}|$: Augmented phrase table size.

Punctuation Insertion

Source language punctuation is modeled by training a 3-gram language model on a punctuated corpus. Punctuation insertion is performed before translation, using the SRILM tool hidden-ngram.

Other Pre-/Postprocessing

We tokenized and lowercased all training data sets. Also, we performed Buckwalter transliteration on all Arabic corpora.

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.

(LA 1) The feature function returns the morphological root(s) of the word.

(LA 2) Still-remaining unknown words go through a second step in which the feature function returns an orthographical normalization of the word obtained by removing all the vowels and diacritics.

Case Restoration

After decoding, target language case information is automatically restored using the **Moses** recasing tool. A lowercase-to-truecase translation model is trained and applied on the translation outputs, together with a few simple rules.

5. Results and Discussion

It is surprising to note that the Chinese-to-Spanish translation with English as the intermediate language (pivot translation) achieves better BLEU scores than the direct translation. We suspect this is due to the similarity of the 2008 test set to the pivot training corpora.

Task	CRR		ASR	
	devset3	test	devset3	test
PIVOT_CES	25.71	**32.94**	20.77	**29.40**
BTEC_CES	**32.40**	29.07	**25.67**	26.85

CH-SP BLEU scores of BTEC vs. PIVOT tasks

6. Conclusion

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.