The QMUL System Description for IWSLT 2010

Sirvan Yahyaei* and Christof Monz**

Queen Mary, University of London* University of Amsterdam**

Introduction

- Arabic-English, French-English and Turkish-English language pairs of the BTEC task
- Our main focus in this submission was on improving the reordering capabilities of the decoder
- improvements were gained by experimenting with different word-alignment strategies and dealing with out of vocabulary (OOV) words.

Preprocessing

- **English:** simple tokenisation and lower-casing.
- **Arabic:** all the diacritics removed, numbers and punctuations normalised, Buckwalter's morphological analyser is used to tokenise.
- **French:** a simple tokeniser which works for all European languages in addition to lower-casing.
- **Turkish:** Morfessor [1] is used for tokenisation. All the words are lower-cased.

Decoder

- Berkeley Aligner [2], for all three language pairs, with 5 joint iterations of IBM model 1, IBM model 2 and HMM.
- phrase translation probabilities and lexical probabilities for both directions.
- 4-gram language model.
- phrase, word and distance-based re-ordering penalties.
- discriminative reordering model

OOV

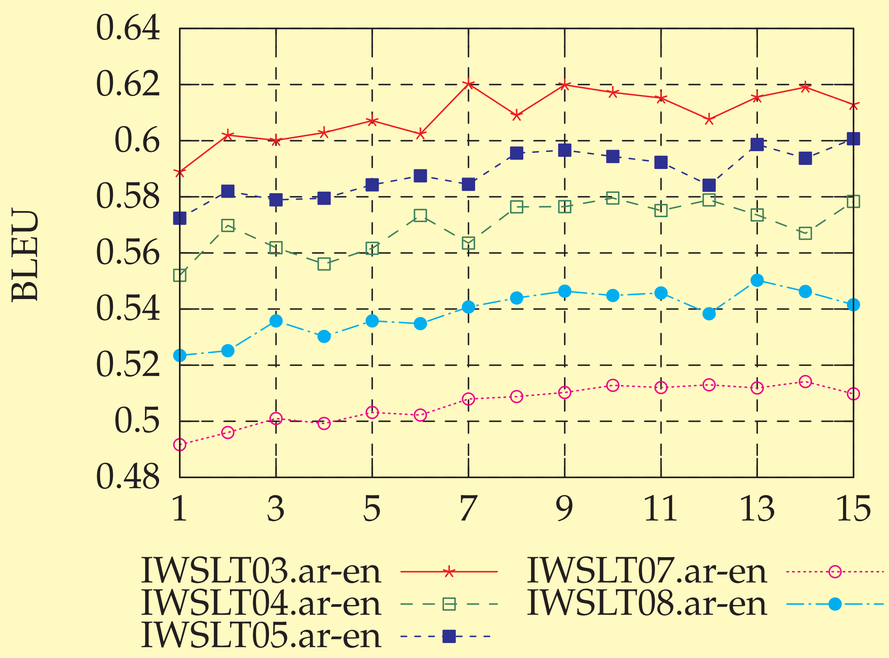
- for a small size training data, unknown words are a significant problem.
- many of the unknown words are morphological variations of known words
- stemming algorithms are used to find matches of the unknown words
- the unknown word is replaced by the unstemmed word matched in the training data

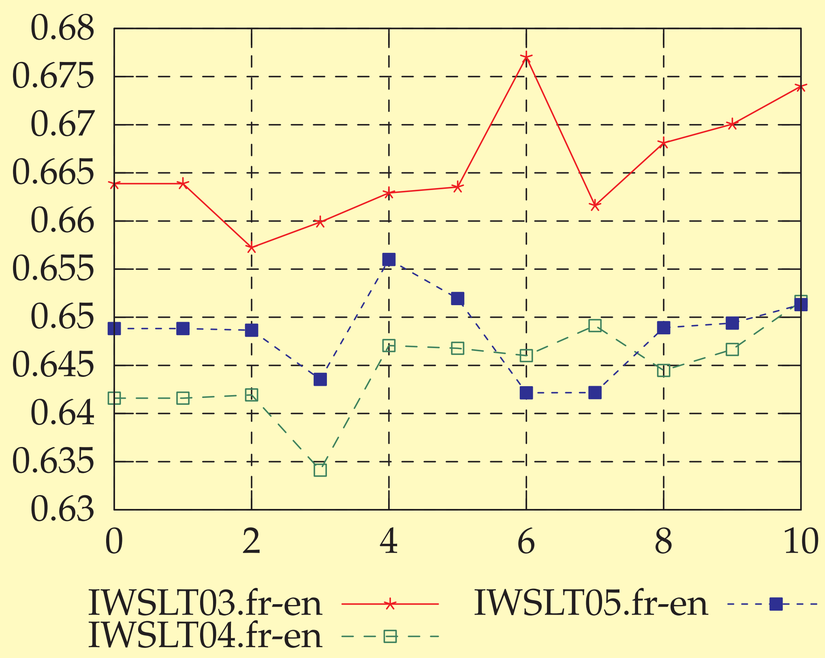
OOV Statistics

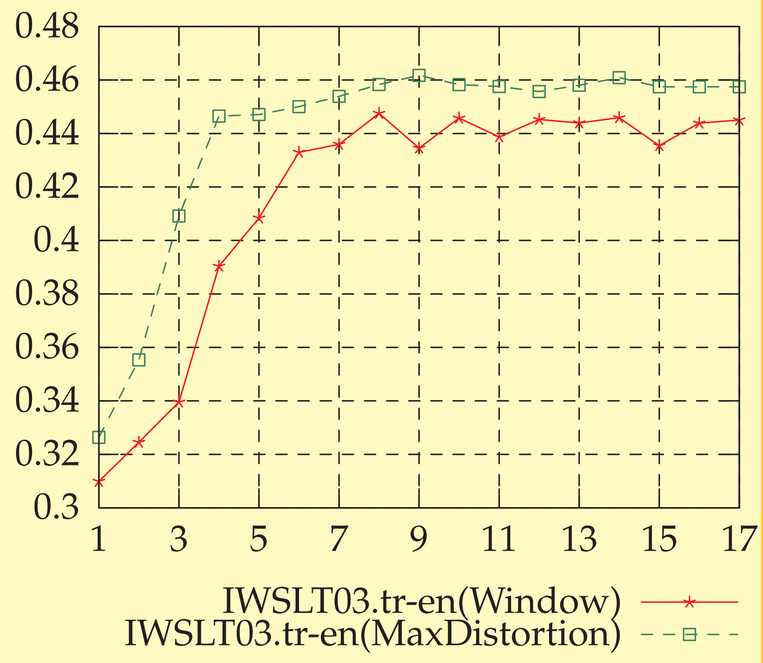
Number of OOV tokens in the test sets before finding replacements and after.

Data set	Source language	Words	Vocabulary	OOV before	OOV after
IWSLT09.ar-en	Arabic	3135	1039	155	82
IWSLT10.ar-en	Arabic	3207	1096	127	54
IWSLT09.fr-en	French	3877	888	70	45
IWSLT10.fr-en	French	3813	901	61	43
IWSLT09.tr-en	Turkish	2944	1071	137	79
IWSLT10.tr-en	Turkish	2910	1102	125	76

Distortion Limit and Constraint







Discriminative Reordering

A maximum entropy model that predicts the length of the next jump based on global and local features. Features for a jump from j to j' in a sentence f_1^J are:

- f_j, $f_{j'}$, $f_j + f_{j'}$
- all the words between j and j'
- part of speech tags of the above words: POS(f_j), POS(f'_j), ...
- bigrams: $f_{j-1} + f_j$ and $f_{j'} + f_{j'+1}$
- bigram part of speech tags of j, j' and the words between them.
- a binary feature indicating that both j and j' are in the same syntactic chunk or not?
- binary feature indicating that f_1^J contains a question mark or not?
- is there a question mark or full stop between j and j'?
- is there a punctuation mark between j and j'?

Dynamic Distortion

Both translation quality and decoding speed are influenced by changing the distortion limit parameter. The discriminative reordering model is used to compute the probability of jumps from that source position to every other position. The jumps after each source position j in the sentence f_1^J is computed as:

$$s_j(j') = \prod_{j''=j}^{j''=j'} p(d_{j,j''}|f_1^J, j, j'') \times \prod_{j''=j'+1}^{j''=J+1} (1 - p(d_{j,j''}|f_1^J, j, j'')) \quad (1)$$

The final distortion limit estimated by this approach for position j equals to:

$$dl(j) = \text{distance}(j, \arg\max_{j'}\{s_j(j')\}) \quad (2)$$

Results

Offcial BLEU results of the primary runs of the QMUL system on BTEC data sets.

Data set	Arabic-English	French-English	Turkish-English
IWSLT09	0.5276	0.6180	0.5354
IWSLT10	0.4425	0.5362	0.5128

References

[1] Mathias Creutz and Krista Lagus. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0 In *Publications in Computer and Information Science, Report A81*, 2005

[2] Percy Liang and Ben Taskar and Dan Klein. Alignment by Agreement. In *Proceedings of HLT-NAACL*, 2006