A Appendices

A.1 Features

We build two sets of features, one for finding the corresponding translation t_j and r_k for each s_i, the other for checking the equivalence if t_j and r_k.

A.1.1 Bilingual alignment features

Bilingual alignment features assume that if s_i is aligned to any existing t_j and r_k by bilingual alignment methods. We rely on Giza++ (Och and Ney, 2000) to find bilingual alignment between S and T (R). However, as Giza++ generates noisy alignment especially for low-frequent words, we propose a bunch of features to complement Giza++ results for more accurate alignments. We apply the same types of features to both T and R, hence only the feature for the alignment between S and T are described here. Letting s_i is aligned to t_j[8].

POS tag: This is a feature on the source. The intuition is that functional words, indicated by POS tags, usually do not need translation. Hence it may not need to align to any words in the target language.

NER feature: This is a binary feature to indicate whether the NER tags of s_i and t_j are the same. A correctly aligned word pair should have the same NER tag. Also this feature helps to determine the `WT` and `MT` error class.

Giza++ confidence: Besides Giza++ translation probability, we also use the word frequency of s_i and t_j in our parallel corpus to penalize the alignment confidence score for low frequency words.

Word-level similarity: We obtain another alternate translation t'_j for s_i using a dictionary and compare the similarity between t_j and t'_j using morphology (e.g., edit distance, number of gram letters overlap, common prefix) and semantic dimension (e.g., word embedding similarity).

Context: We assume that s_i and t_j should be aligned if they are translation equivalent and the same applies to the words linking to s_i and t_j in the dependency tree parsed by Standford Corenlp (Manning et al., 2014). Thus, we define context as the number of aligned-pairs among words linking to s_i and t_j.

Sentence-level translation quality feature: This include the sentence-level QE shared-task baseline features used in (Specia et al., 2018). Such features are to estimate the overall quality of the MT hypothesis. The intuition is the that better sentence level translation the less probable word-level misalignment.

A.1.2 Monolingual equivalence checking

We leverage on monolingual alignment to compare t_j and r_k and expect that semantic equivalent words can be aligned by monolingual alignment methods. We use the state-of-the-art alignment tool proposed by Sultan et al. (2014). The tool leverages on paraphrase lexicon (Pavlick et al., 2015) and dependency relations to find equivalent expression between two sentences in the same language. The following features are used.

Monolingual alignment feature: A binary feature to indicate whether t_j and r_k are aligned by tool proposed by (Sultan et al., 2014). This feature is set to be 0 if either t_j or r_k does not exist.

NER feature: A binary feature to indicate whether the NER tag of t_j and r_k are the same.

[8] Following Hu et al. (2018), we make adjustment based on by penalizing the frequency of both words.