CS671: NLP
Neural Network based translation and parallel corpus generation

Basis: As of the current scheme in machine translation, Statistical Machine Translation (SMT) is prefered to Neural Network based systems. Also, making a system that learns translation requires the availability of a one to one correspondence between the sources and target sentences, i.e., having parallel corpora at one's disposal is crucial. However, commonly available parallel corpora contain hardly more than a 100,000 words. Therefore, the translators trained using them are naturally weak.

Project schema: We propose to train a system that learns translation as well as generates parallel corpus using comparable corpus (which is readily available) for any 2 languages. We achieve this by:

1. Training a weak translator using the limited parallel corpus available.
2. Assuming we have a corpus X and its comparable counterpart Y, we use this weak translator to translate X into Y's language yielding a corpus Z.
3. An aligner like Hunalign or LF Aligner (again based on hunalign) is used to match the concepts within sentences in Z to the concepts within sentences in Y.
4. The above step outputs matching pairs of sentences in Y and Z (both in the same language, of course). For instance, if Y had sentences from $Y_1, Y_2, \ldots, Y_N$ while Z had sentences from $Z_1, Z_2, \ldots, Z_M$, the aligner produces sentence pairs: $\{Y_1, Z_1\}, \{Y_2, Z_2\}, \ldots, \{Y_K, Z_K\}$. Note that the output numbering may not be the same as input numbering of sentences.
5. The Z sentences in the pair are mapped back to their counterparts in X and we get pairs $\{Y_1, X_1\}, \{Y_2, X_2\}, \ldots, \{Y_K, X_K\}$.
6. Note that the above generated parallel corpus is free of any noise associated with translation.
7. The weak translator is retrained on the generated parallel corpus in a similar way.

Prefered languages are English and Hindi.

Papers referenced

1. *Kalchbrenner and Blunsom*, **Recurrent Continuous Translation Models**, 2013
2. *Sutskever et al*, **Sequence to Sequence Learning with Neural Networks**, 2014
3. *Cho et al*, **On the Properties of Neural Machine Translation: Encoder–Decoder Approaches**, 2014
4. *Hermann and Blunsom*, **Multilingual Distributed Representations without Word Alignment**, 2014
5. *Cho et al*, **Neural Machine Translation by jointly learning to align and translate**, ICLR 2015
6. *Wolk and Marasek*, **Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs,** 2014