Improving Word Alignment with Language Model Based Confidence Scores
Nguyen Bach, Qin Gao, and Stephan Vogel

Sentence Pair Probability

In IBM word alignment models, re-estimating the model parameters depends on the empirical probability \( P(e_k, f_k) \) for each sentence pair \((e_k, f_k)\). During the EM training, all counts of events, e.g., word pair counts, distortion model counts, etc., are weighted by \( P(e_k, f_k) \). For example, in IBM Model 1 the lexicon probability of source word \( e \) given target word \( f \) is calculated as:

\[
p(f|e) = \frac{\sum:\lambda \sum_{i} c_{i}(f|e, f_{i})}{\sum_{i} \sum_{k} c_{k}(f|e, f_{k})}
\]

\[
p(f|e, f_{k}) = \frac{\sum_{i} \sum_{k} \hat{p}(e_{i}|e, f_{k}) \sum_{a} p(a|e_{a}, f_{k}) \sum_{j} \delta(f, f_{j}) \delta(e, e_{a})}{\sum_{i} \sum_{k} \sum_{a} \hat{p}(e_{i}|e, f_{k}) \sum_{j} \delta(f, f_{j}) \delta(e, e_{a})}
\]

\( \hat{p}(e_k, f_k) \) determines how much the alignments of sentence pair \((e^k, f^k)\) contribute to the model parameters. \( \hat{p}(e_k, f_k) \) is estimated by MLE on the full sentence pairs of training data.

Motivation

- It’s helpful if \( \hat{p}(e_k, f_k) \) can approximate true distribution \( P(e_k, f_k) \).
- MLE is valid when training data is infinite. However, the assumption is invalid if the data source is finite. In the training corpora, most sentences occur only one time, and thus \( \hat{p}(e_k, f_k) \) will be uniform.
- \( \hat{p}(e_k, f_k) \) can be seen as prior of models. Some sentences could be more valuable, reliable, appropriate, and should therefore have a higher weight in the training.

Proposed Approach

\( \hat{p}(e^k, f^k) \sim \text{sentence pair confidence (sc):} \) Quality of the sentence pair for training alignment models; use general language models in both source and target to compute.

\[
\hat{P}(e^k, f^k) = \text{genre-dependent sentence pair confidence (gcds):} \text{ Adopt training data toward a target genre. Use genre-dependent language models to assign sentence pair confidence.}
\]

\[ g_{gcds}(e^k, f^k) = s_c(e^k, f^k | g) \quad (5) \]

Sentence-dependent phrase alignment confidence (sdpc): given a phrase pair \((e_p, f_p)\), track from which sentence pairs the phrase pair was extracted; add a feature in phrase pairs

\[
s_{dp}(e_{p}, f_{p}) = \exp \left( \sum_{a} p(a | e_{a}, f_{p}) \log \frac{\hat{p}(e_{a} | e_{p}, f_{p})}{\hat{p}(e_{a})} \right)
\]

\[
S(e_{p}, f_{p}) = \exp \left( \sum_{a} p(a | e_{a}, f_{p}) \log \frac{\hat{p}(e_{a} | e_{p}, f_{p})}{\hat{p}(e_{a})} \right)
\]

Experimental Results

EN → ES; training & test data from 2 genres Europarl and News-Commentary; Moses, SRILM, multi-threaded GIZA+++

Calculated gcds for Europal and News-Commentary training data using NC, EP, and NC+EP(NE) LMs. For each sentence we computed the difference of gcds between NC and EP LM, namely \( gcds_{nc} \rightarrow gcds_{ep} \), and plot histogram. Similar analysis have been perform on NC-NE and NE-EP.

General Conclusion

- Weight sentence pairs by LMs is better than weight by MLE.
- Improvements are obtained by using sentence pair confidence scores; using EP LM gains best scores.
- No evidence to show that using genre-dependent sentence pair confidence (gcds) will provide better result comparing with general confidence. Test set model perplexities drop by using gcds, but translation results are going against expectation.
- Did not observe consistent improvements by using sentence-dependent phrase alignment confidence.