Improving Word Alignment with Language Model Based Confidence Scores

Nguyen Bach, Qin Gao, and Stephan Vogel

interACT

Sentence Pair Probability

In IBM word alignment models, re-estimating the model parameters depends on the **empirical probability** $\hat{P}(e^k, f^k)$ 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 $\hat{P}(e^k, f^k)$. For example, in IBM Model 1 the lexicon probability of source word **f** given target word **e** is calculated as :

$$p(\mathbf{f}|\mathbf{e}) = \frac{\sum_k c(\mathbf{f}|\mathbf{e}; e^k, f^k)}{\sum_{k,\mathbf{f}} c(\mathbf{f}|\mathbf{e}; e^k, f^k)} \quad (1)$$

$$c(\mathbf{f}|\mathbf{e}; e^k, f^k) = \sum_{e^k, f^k} \hat{\mathrm{P}}(\mathbf{e}^k, \mathbf{f}^k) \sum_a P(a|e^k, f^k) \sum_j \delta(\mathbf{f}, f_j^k)\delta(\mathbf{e}, e_{a_j}^k) \quad (2)$$

$\hat{P}(e^k, f^k)$ **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

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

$$\begin{aligned} \mathcal{L}(e^k) &= \frac{1}{|e^k|} \textstyle\sum_{e_i^k \in e^k} \log P(e_i^k|h) \\ \mathcal{L}(f^k) &= \frac{1}{|f^k|} \textstyle\sum_{f_j^k \in f^k} \log P(f_j^k|h) \quad (3) \\ \mathcal{L}(e^k, f^k) &= [\mathcal{L}(e^k) + \mathcal{L}(f^k)]/2 \end{aligned}$$

$$sc(e^k, f^k) = \exp(\mathcal{L}(e^k, f^k)). \quad (4)$$

$\hat{P}(e^k, f^k)$ ~ **genre-dependent sentence pair confidence (*gdsc*)**: Adopt training data toward a target genre. Use genre-dependent language models to assign sentence pair confidence.

$$gdsc(e^k, f^k) = sc(e^k, f^k|g) \quad (5)$$

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

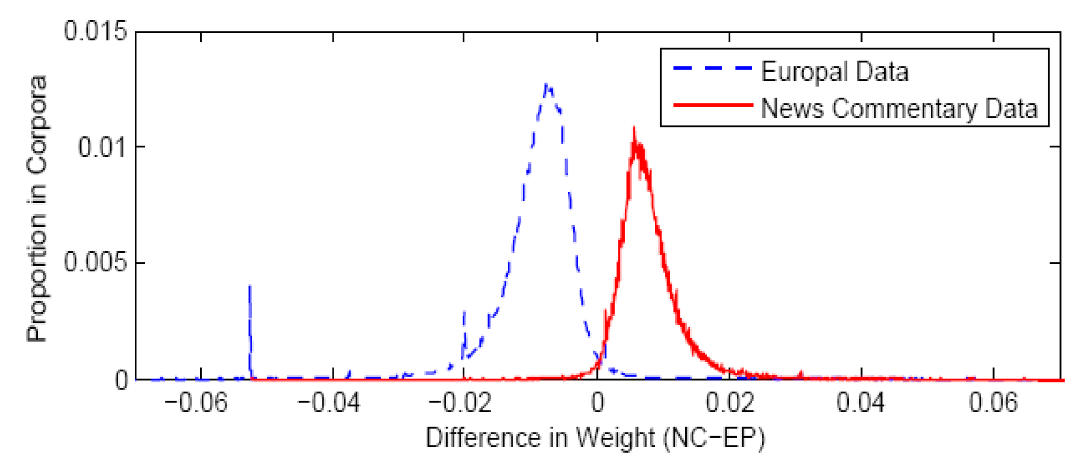
$$\begin{aligned} sdpc(ep, fp) &= \exp \frac{\sum_{(e^k, f^k) \in \mathcal{S}(ep, fp)} \log sc(e^k, f^k)}{|\mathcal{S}(ep, fp)|} \\ \mathcal{S}(ep, fp) &= \{(e^k, f^k) | ep \in e^k, fp \in f^k\} \quad (6) \end{aligned}$$

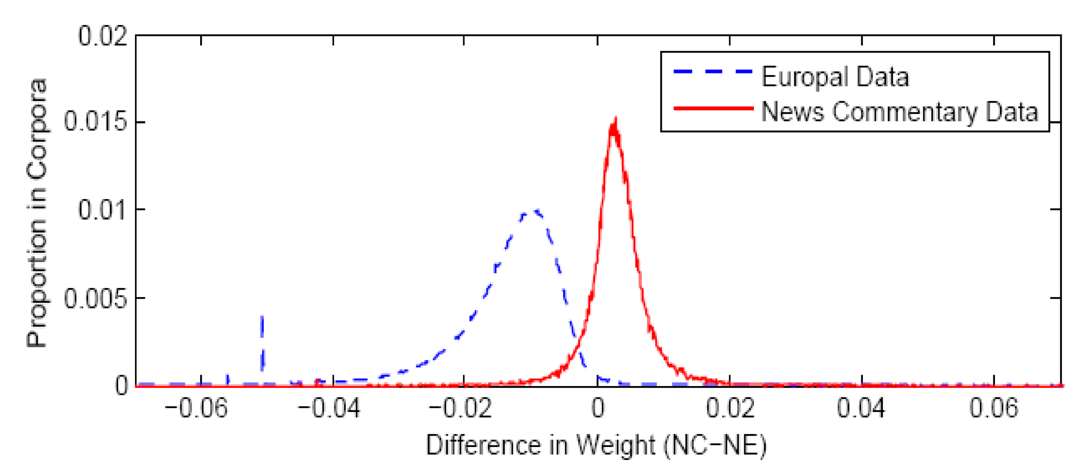
Experimental Results

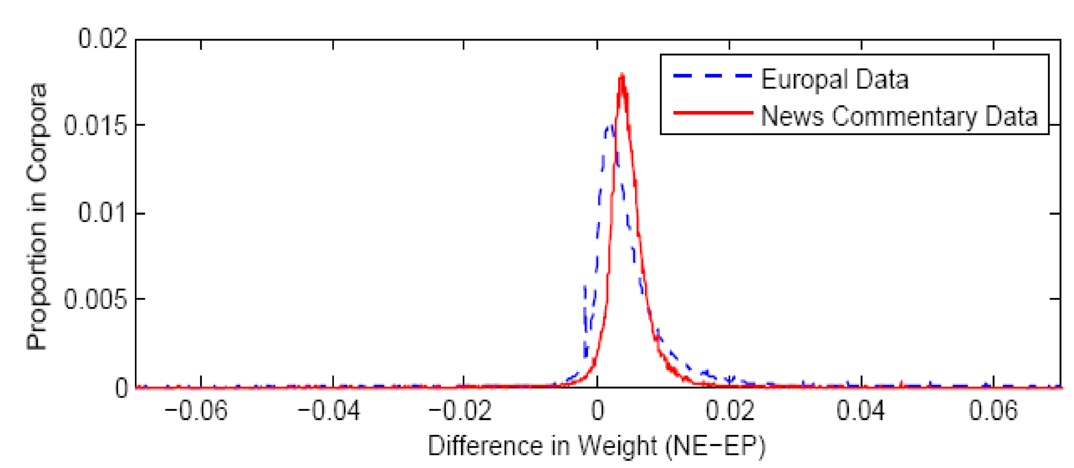
EN ↔ ES; training & test data from 2 genres Europarl and News-Commentary; Moses, SRILM, multi-threaded GIZA++.

	English	Spanish
Europarl (E)		
sentence pairs	1,258,778	
unique sent. pairs	1,235,134	
avg. sentence length	27.9	29.0
# words	35.14 M	36.54 M
vocabulary	108.7 K	164.8 K
News-Commentary (NC)		
sentence pairs	64,308	
unique sent. pairs	64,205	
avg. sentence length	24.0	27.4
# words	1.54 M	1.76 M
vocabulary	44.2 K	56.9 K

Histogram of weight differences







Calculated ***gdsc*** for Europal and News-Commentary training data using NC, EP and NC+EP(NE) LMs.

For each sentence we computed the difference of *gdsc* between NC and EP LM, namely ***$gdsc_{NC}$ - $gdsc_{EP}$***, and plot histogram.

Similar analysis have been perform on NC-NE and NE-EP.

Model 4 train perplexities when using Sentence Pair Confidence scores

		None	EP+ NC	NC	EP
Train	En → Es	46.76	**42.36**	42.97	44.47
	Es → En	70.18	**62.81**	62.95	65.86
Test	EP (En → Es)	91.13	90.89	91.84	**90.77**
	NC (En → Es)	53.04	53.44	**51.09**	55.94
	EP (Es → En)	126.56	125.96	123.23	**122.11**
	NC (Es → En)	81.39	81.28	**78.23**	80.33

Perplexities drop significantly in training data of two translation directions, and in test sets, perplexities also drop in genre.

Performance of sentence pair confidence scores (*sc* , *gdsc*) in BLEU

	E06	E07	NCd	NCt1	NCt2
			EN → EN		
None	33.26	33.23	36.06	35.56	35.64
NC+EP	33.23	32.29	36.12	35.47	35.97
NC	**33.43**	**33.39**	36.14	35.27	35.68
EP	33.36	**33.39**	**36.16**	**35.63**	**36.17**
			EN → ES		
None	**33.33**	32.25	35.1	34.08	34.43
NC+EP	33.23	**32.29**	**35.12**	34.56	34.89
NC	33.3	32.27	34.91	34.07	34.29
EP	33.08	**32.29**	35.05	**34.52**	**35.03**

The improvements on NC sets are obvious, especially on held-out evaluation sets NC_t & NC_{t1}; using EP obtained the best performance.

Performance of sentence-dependent phrase alignment confidence (sdpc)

	E06	E07	NCd	NCt1	NCt2
			ES → EN		
None	33.26	33.23	36.06	35.56	35.64
NC+EP +*sdpc*	**33.54**	**33.39**	36.07	35.38	35.85
NC +*sdpc*	33.17	33.31	35.96	**35.74**	36.04
EP +*sdpc*	33.44	32.87	**36.22**	35.63	**36.09**
			EN → ES		
None	**33.33**	32.25	**35.1**	34.08	34.43
NC+EP +*sdpc*	33.28	32.45	34.82	33.68	33.86
NC +*sdpc*	33.13	**32.47**	34.01	**34.34**	**34.98**
EP +*sdpc*	32.97	32.2	34.26	33.99	34.34

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 (gdsc) will provide better result comparing with general confidence. Test set model perplexities drop by using gdsc, but translation results are going against expectation.
- Did not observe consistent improvements by using sentence-dependent phrase alignment confidence .

Carnegie Mellon

lti