SIMULATING SENTENCE PAIRS SAMPLING VIA SOURCE AND TARGET LANGUAGE MODELS

Nguyen Bach

Joint work with Qin Gao and Stephan Vogel

Outline

- Motivations
- Proposed Approach
- Experiments

Motivations

Weighting Sentence Pairs

- Normal word alignment
  - Each sentence pair (e^k,f^k) is assigned an empirical probability $\hat{P}(e^k, f^k)$
  - IBM Model 1: lexicon probability of source word **f** given target **e**

$$
\begin{aligned}
p(\mathbf{f}|\mathbf{e}) &= \frac{\sum_k c(\mathbf{f}|\mathbf{e}; e^k, f^k)}{\sum_{k,\mathbf{f}} c(\mathbf{f}|\mathbf{e}; e^k, f^k)} && (1) \\
c(\mathbf{f}|\mathbf{e}; e^k, f^k) &= \sum_{e^k, f^k} \hat{P}(e^k, f^k) \sum_a P(a|e^k, f^k) \cdot && (2) \\
& \sum_j \delta(\mathbf{f}, f_j^k)\delta(\mathbf{e}, e_{a_j}^k)
\end{aligned}
$$

- $\hat{P}(e^k, f^k)$ is estimated by MLE on the full sentence pairs which would give most uniform probabilities (~ 1/S)

Motivation

- It's helpful if $\hat{P}(e^k, f^k)$ can approximate the true distribution $P(e^k, f^k)$
- $\hat{P}(e^k, f^k)$ is a prior
- Some sentences could be more valuable, reliable, appropriate, and should therefore have a **higher weight** in the training

- Can we have better a approximation for $\hat{P}(e^k, f^k)$?

Proposed Approach

Proposed approach

- $\hat{P}(e^k, f^k) \sim$ sentence pair confidence (***sc***)
  - Quality of sentence pair for training the alignment model
- $\hat{P}(e^k, f^k) \sim$ genre-dependent sentence pair confidence (***gdsc***)
  - Appropriateness of a sentence pair to train a system for a specific genre
- Sentence-dependent phrase alignment confidence (***sdpc***) scores
  - Which sentence pairs the phrase pair was extracted

Sentence pair confidence (sc)

- It's hard to compute $P(e^k, f^k)$ without knowing $P(e^k|f^k)$ which is estimated during the alignment process
- Assumption

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

- $P(e^k)$, $P(f^k)$ can be estimated by source and target language models

Sentence pair confidence (sc)

- Average log likelihood of each sentence pair

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

- Sentence pair confidence score (***sc***)

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

$$= \sqrt{\left(\prod_{e_i^k \in e^k} P(e_i^k|h)\right)^{-|e^k|} \left(\prod_{f_j^k \in f^k} P(f_j^k|h)\right)^{-|f^k|}} \quad (4)$$

Genre-dependent sentence pair confidence (gdsc)

- Adopt training data toward a target genre.
- Use genre-dependent language models to assign sentence pair confidence
- Given genre ***g***

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

- Average likelihood of each sentence is estimated by genre-specific language models

Sentence-dependent phrase alignment confidence (sdpc)

- We want to put ***sc*** into decoding process
  - Add a feature in phrase pairs
- Track from which sentence pairs the phrase pair was extracted
- Given a phrase pair (*ep,fp*), the ***sdpc*** score

$$sdpc(ep, fp) = \exp \frac{\sum_{(e^k, f^k) \in S(ep,fp)} \log sc(e^k, f^k)}{|S(ep, fp)|}$$

$$S(ep, fp) = \{(e^k, f^k) | ep \in e^k, fp \in f^k\} \qquad (6)$$

where *S(ep, fp)* is the set of sentences that the phrase pair come from

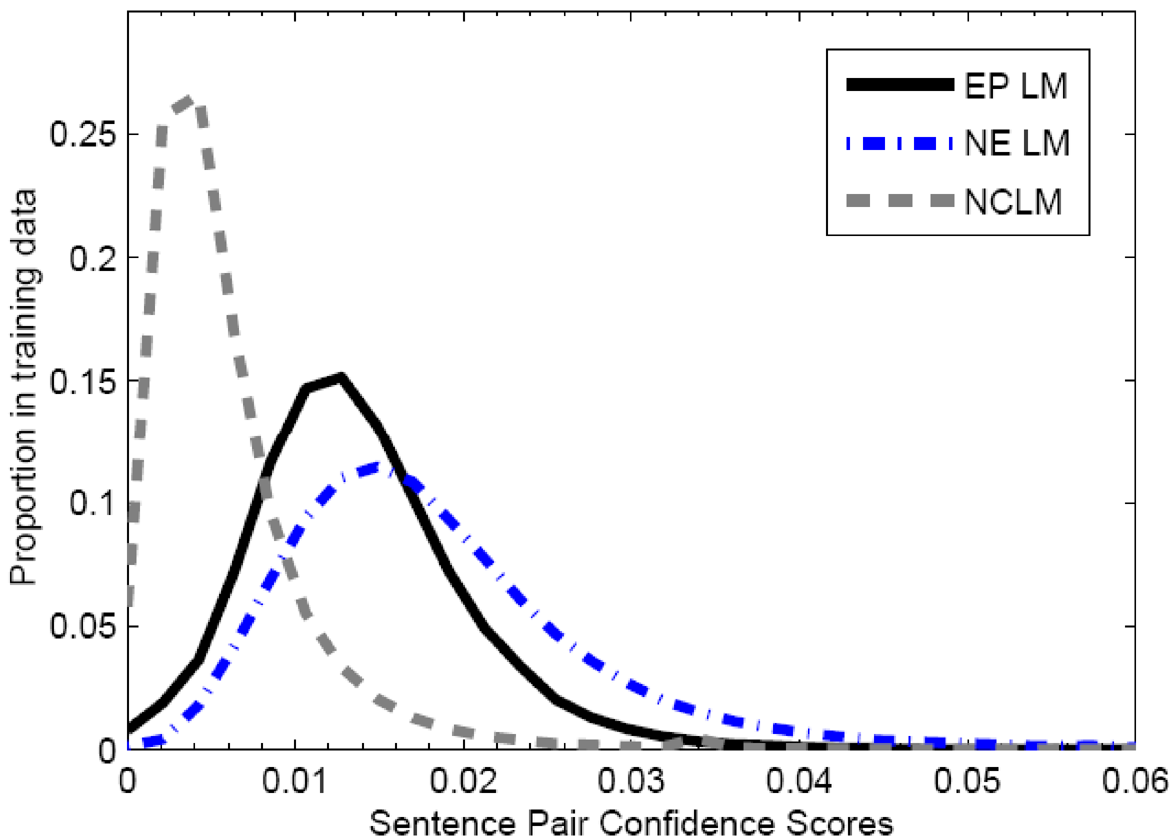
Experimental Results

Set-up

- EN ↔ ES
- Training & test data from 2 genres Europarl and News-Commentary (ACL'08-WMT)
- Standard toolkits Moses, SRILM, GIZA++ (multi-threaded)

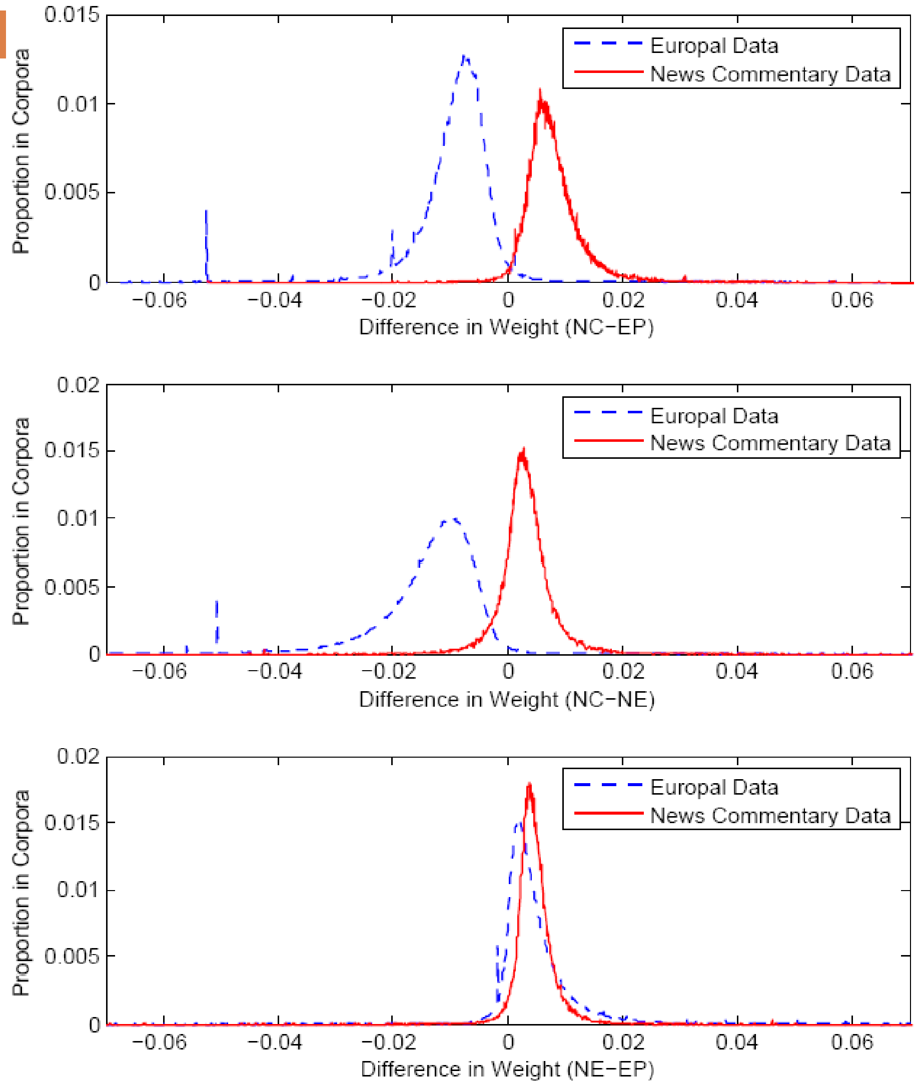
	English	Spanish
Europarl (E)		
sentence pairs	1,258,778	
unique sent. pairs	1,235,134	
avg. sentence length	27.9	29.0
# words	35.14 M	36.54 M
vocabulary	108.7 K	164.8 K
News-Commentary (NC)		
sentence pairs	64,308	
unique sent. pairs	64,205	
avg. sentence length	24.0	27.4
# words	1.54 M	1.76 M
vocabulary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

Histogram of *sc* weights



- Calculated *sc* for the whole training data using NC, EP and NC+EP(NE) LMs.
- Many sentences get a much higher score in training than using MLE

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 $\boldsymbol{gdsc_{NC}}$ - $\boldsymbol{gdsc_{EP}}$, and plot histogram.
- Similar analysis have been perform on NC-NE and NE-EP.

IBM Model 4 train perplexities when using Sentence Pair Confidence scores

- IBM Model-4 train perplexities on train and test data

		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.
- In test sets, perplexities also drop in genre which implied a better word alignment model had been learned.

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

	E06	E07	NCd	NCt1	NCt2
			ES → 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 News-Commentary sets are obvious, especially on held-out evaluation sets NCt and NCt1; using EP obtained the best performance
- No evidence to show that using genre-dependent confidence will provide better result comparing with general confidence.

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

- Across development and held-out sets the gains from *sdpc* are inconsistent

Conclusion

- We developed
  - sentence pair confidence (***sc***)
  - *genre-dependent* sentence pair confidence (***gdsc***)
  - *sentence-dependent* phrase alignement confidence (***sdpc***) *scores.*
- Using source and target language models to estimate scores.
- Experimental results shown that
  - Better approximation for empirical probability of sentence pairs. Improvements are obtained by using sentence pair confidence scores; using EP LM gain best scores.
  - No evidence to show that using *gdsc* will provide better result comparing with general confidence.
  - Test set model perplexities drop by using ***gsdc*,** but translation results are going against expectation
  - Did not observe consistent improvements by using *sdpc*

THANK YOU