Improving Word Alignment with Language Model Based Confidence Scores

Nguyen Bach, Qin Gao, Stephan Vogel
InterACT, Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{nbach, qing, vogel+}@cs.cmu.edu

Abstract

This paper describes the statistical machine translation systems submitted to the ACL-WMT 2008 shared translation task. Systems were submitted for two translation directions: English→Spanish and Spanish→English. Using sentence pair confidence scores estimated with source and target language models, improvements are observed on the News-Commentary test sets. Genre-dependent sentence pair confidence score and integration of sentence pair confidence score into phrase table are also investigated.

1 Introduction

Word alignment models are a crucial component in statistical machine translation systems. When estimating the parameters of the word alignment models, the sentence pair probability is an important factor in the objective function and is approximated by the empirical probability. The empirical probability for each sentence pair is estimated by maximum likelihood estimation over the training data (Brown et al., 1993). Due to the limitation of training data, most sentence pairs occur only once, which makes the empirical probability almost uniform. This is a rather weak approximation of the true distribution.

In this paper, we investigate the methods of weighting sentence pairs using language models, and extended the general weighting method to genre-dependent weight. A method of integrating the weight directly into the phrase table is also explored.

2 The Baseline Phrase-Based MT System

The ACL-WMT08 organizers provided Europarl and News-Commentary parallel corpora for English ↔ Spanish. Detailed corpus statistics is given in Table 1. Following the guidelines of the workshop we built baseline systems, using the lower-cased Europarl parallel corpus (restricting sentence length to 40 words), GIZA++ (Och and Ney, 2003), Moses (Koehn et al., 2007), and the SRI LM toolkit (Stolcke, 2002) to build 5-gram LMs. Since no News development sets were available we chose News-Commentary sets as replacements. We used test-2006 (E06) and nc-devtest2007 (NCd) as development sets for Europarl and News-Commentary; test-2007 (E07) and nc-test2007 (NCt) as held-out evaluation sets.

	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

Table 1: Statistics of English↔Spanish Europarl and News-Commentary corpora

To improve the baseline performance we trained systems on all true-cased training data with sentence length up to 100. We used two language models, a 5-gram LM build from the Europarl corpus and a 3-gram LM build from the News-Commentary data. Instead of interpolating the two language models, we explicitly used them in the decoder and optimized their weights via minimum-error-rate (MER) training (Och, 2003). To shorten the training time, a multi-threaded GIZA++ version was used to utilize multi-processor servers (Gao and Vogel, 2008). Other parameters were the same as the baseline system. Table 2 shows results in lowercase BLEU (Papineni et al., 2002) for both the baseline (B) and the improved baseline systems (B5) on development and held-

Proceedings of the Third Workshop on Statistical Machine Translation, pages 151–154,
Columbus, June 2008.

out evaluation sets. We observed significant gains for the News-Commentary test sets. Our improved baseline systems obtained a comparable performance with the best English↔Spanish systems in 2007 (Callison-Burch et al., 2007).

Pairs		Europarl		NC	
		E06	E07	NCd	NCt
En→Es	B	33.00	32.21	31.84	30.56
	B5	33.33	32.25	35.10	34.08
Es→En	B	33.08	33.23	31.18	31.34
	B5	33.26	33.23	36.06	35.56

Table 2: NIST-BLEU scores of baseline and improved baseline systems experiments on English↔Spanish

3 Weighting Sentence Pairs

3.1 Problem Definition

The quality of word alignment is crucial for the performance of the machine translation system.

In the well-known so-called IBM word alignment models (Brown et al., 1993), 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 $\mathbf{f}$ given target word $\mathbf{e}$ is calculated as (Och and Ney, 2003):

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

Therefore, the distribution of $\hat{P}(e^k, f^k)$ will affect the alignment results. In Eqn. 2, $\hat{P}(e^k, f^k)$ determines how much the alignments of sentence pair (e^k, f^k) contribute to the model parameters. It will be helpful if the $\hat{P}(e^k, f^k)$ can approximate the true distribution of $P(e^k, f^k)$.

Consider that we are drawing sentence pairs from a given data source, and each *unique* sentence pair (e^k, f^k) has a probability $P(e^k, f^k)$ to be observed. If the training corpora size is infinite, the normalized frequency of each unique sentence pair will converge to $P(e^k, f^k)$. In that case, equally assigning a number to each occurrence of (e^k, f^k) and normalizing it will be valid. However, the assumption is invalid if the data source is finite. As we can observe in the training corpora, most sentences occur only one time, and thus $\hat{P}(e^k, f^k)$ will be uniform.

To get a more informative $\hat{P}(e^k, f^k)$, we explored methods of weighting sentence pairs. We investigated three sets of features: sentence pair confidence (sc), genre-dependent sentence pair confidence ($gdsc$) and phrase alignment confidence (pc) scores. These features were calculated over an entire training corpus and could be easily integrated into the phrase-based machine translation system.

3.2 Sentence Pair Confidence

We can hardly compute the joint probability of $P(e^k, f^k)$ without knowing the conditional probability $P(e^k|f^k)$ which is estimated during the alignment process. Therefore, to estimate $P(e^k, f^k)$ before alignment, we make an assumption that $\hat{P}(e^k, f^k) = P(e^k)P(f^k)$, which means the two sides of sentence pair are independent of each other. $P(e^k)$ and $P(f^k)$ can be obtained by using language models. $P(e^k)$ or $P(f^k)$, however, can be small when the sentence is long. Consequently, long sentence pairs will be assigned low scores and have negligible effect on the training process. Given limited training data, ignoring these long sentences may hurt the alignment result. To compensate this, we normalize the probability by the sentence length. We propose the following method to weighting sentence pairs in the corpora. We trained language models for source and target language, and the average log likelihood (AVG-LL) of each sentence pair was calculated by applying the corresponding language model. For each sentence pair (e^k, f^k), the AVG-LL $\mathcal{L}(e^k, f^k)$ is

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

where $P(e_i^k|h)$ and $P(f_j^k|h)$ are ngram probabilities. The sentence pair confidence score is then given by:

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

3.3 Genre-Dependent Sentence Pair Confidence

Genre adaptation is one of the major challenges in statistical machine translation since translation models suffer from data sparseness (Koehn and Schroeder, 2007). To overcome these problems previous works have focused on explicitly modeling topics and on using multiple language and translation models. Using a mixture of topic-dependent Viterbi alignments was proposed in (Civera and Juan, 2007). Language and translation model adaptation to Europarl and News-Commentary have been explored in (Paulik et al., 2007).

Given the sentence pair weighting method, it is possible to adopt genre-specific language models into the weighting process. The genre-dependent sentence pair confidence $gdsc$ simulates weighting the training sentences again from different data sources, thus, given genre g, it can be formulated as:

$$gdsc(e^k, f^k) = sc(e^k, f^k|g) \tag{5}$$

where $P(e_i^k|h)$ and $P(f_j^k|h)$ are estimated by genre-specific language models.

The score generally represents the likelihood of the sentence pair to be in a specific genre. Thus, if both sides of the sentence pair show a high probability according to the genre-specific language models, alignments in the pair should be more possible to occur in that particular domain, and put more weight may contribute to a better alignment for that genre.

3.4 Phrase Alignment Confidence

So far the confidence scores are used only in the training of the word alignment models. Tracking from which sentence pairs each phrase pair was extracted, we can use the sentence level confidence scores to assign confidence scores to the phrase pairs. Let $S(\tilde{e}, \tilde{f})$ denote the set of sentences pairs from which the phrase pair $(\tilde{e}, \tilde{f})$ was extracted. We calculate then a phrase alignment confidence score pc as:

$$pc(\tilde{e}, \tilde{f}) = \exp \frac{\sum_{(e^k, f^k) \in \mathcal{S}(\tilde{e}, \tilde{f})} \log sc(e^k, f^k)}{|S(\tilde{e}, \tilde{f})|} \tag{6}$$

This score is used as an additional feature of the phrase pair. The feature weight is estimated in MER training.

4 Experimental Results

The first step in validating the proposed approach was to check if the different language models do assign different weights to the sentence pairs in the training corpora. Using the different language models NC (News-Commentary), EP (Europarl), NC+EP (both NC and EP) the genre-specific sentence pair confidence scores were calculated. Figure 1 shows the distributions of the differences in these scores across the two corpora. As expected, the language model build from the NC corpus assigns - on average - higher weights to sentence pairs in the NC corpus and lower weights to sentence pairs in the EP corpus (Figure 1a). The opposite is true for the EP LM. When comparing the scores calculated from the NC LM and the combined NC+EP LM we still see a clear separation (Figure 1b). No marked difference can be seen between using the EP LM and the NC+EP LM (Figure 1c), which again is expected, as the NC corpus is very small compared to the EP corpus.

The next step was to retrain the word alignment models using sentences weights according to the various confidence scores. Table 3 shows training and test set perplexities for IBM model 4 for both training directions. Not only do we see a drop in training set perplexities, but also in test set perplexities. Using the genre specific confidence scores leads to lower perplexities on the corresponding test set, which means that using the proposed method does lead to small, but consistent adjustments in the alignment models.

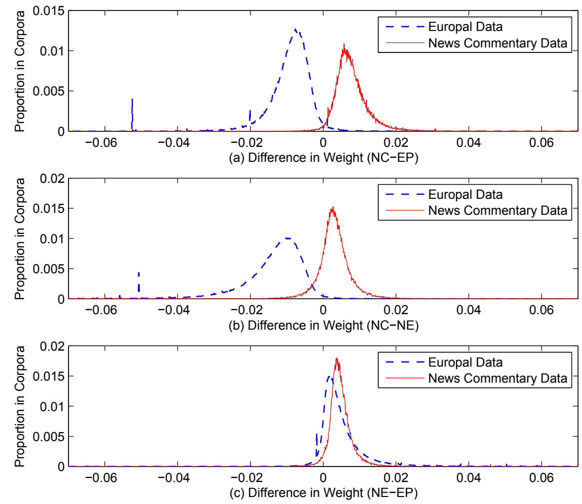

Figure 1: Histogram of weight differences genre specific confidence scores on NC and EP training corpora

		Uniform	NC+EP	NC	EP
train	En→Es	46.76	**42.36**	42.97	44.47
	Es→En	70.18	**62.81**	62.95	65.86
test	NC(En→Es)	53.04	53.44	**51.09**	55.94
	EP(En→Es)	91.13	90.89	91.84	**90.77**
	NC(Es→En)	81.39	81.28	**78.23**	80.33
	EP(Es→En)	126.56	125.96	123.23	**122.11**

Table 3: IBM model 4 training and test set perplexities using genre specific sentence pair confidence scores.

In the final step the specific alignment models were used to generate various phrase tables, which were then used in translation experiments. Results are shown in Table 4. We report lower-cased Bleu scores. We used nc-dev2007 (NCt1) as an additional held-out evaluation set. Bold cells indicate highest scores.

As we can see from the results, improvements are obtained by using sentence pair confidence scores. Using confidence scores calculated from the EP LM gave overall the best performance. While we observe only a small improvement on Europarl sets, improvements on News-Commentary sets are more pronounced, especially on held-out evaluation sets NCt and NCt1. The experiments do not give evidence that genre-dependent confidence can improve over using the general confidence

	Test Set				
	E06	E07	NCd	NCt	NCt1
			Es→En		
B5	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		
B5	**33.33**	32.25	35.10	34.08	34.43
NC+EP	33.23	**32.29**	**35.12**	34.56	34.89
NC	33.30	32.27	34.91	34.07	34.29
EP	33.08	**32.29**	35.05	**34.52**	**35.03**

Table 4: Translation results (NIST-BLEU) using $gdsc$ with different genre-specific language models for Es↔En systems

score. As the News-Commentary language model was trained on a very small amount of data further work is required to study this in more detail.

	Test Set				
	E06	E07	NCd	NCt	NCt1
			Es→En		
B5	33.26	33.23	36.06	35.56	35.64
NC+EP+pc	**33.54**	**33.39**	36.07	35.38	35.85
NC+pc	33.17	33.31	35.96	**35.74**	36.04
EP+pc	33.44	32.87	**36.22**	35.63	**36.09**
			En→Es		
B5	**33.33**	32.25	**35.10**	34.08	34.43
NC+EP+pc	33.28	32.45	34.82	33.68	33.86
NC+pc	33.13	**32.47**	34.01	**34.34**	**34.98**
EP+pc	32.97	32.20	34.26	33.99	34.34

Table 5: Translation results (NIST-BLEU) using pc with different genre-specific language models for Es↔En systems

Table 5 shows experiments results in NIST-BLEU using pc score as an additional feature on phrase tables in Es↔En systems. We observed that across development and held-out sets the gains from pc are inconsistent, therefore our submissions are selected from the B5+EP system.

5 Conclusion

In the ACL-WMT 2008, our major innovations are methods to estimate sentence pair confidence via language models. We proposed to use source and target language models to weight the sentence pairs. We developed sentence pair confidence (sc), genre-dependent sentence pair confidence ($gdsc$) and phrase alignment confidence (pc) scores. Our experimental results shown that we had a better word alignment and translation performance by using $gdsc$. We did not observe consistent improvements by using phrase pair confidence scores in our systems.

Acknowledgments

This work is in part supported by the US DARPA under the GALE program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, volume 19(2), pages 263–331.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proc. of the ACL 2007 Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical translation with mixture modelling. In *Proc. of the ACL 2007 Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Proc. of the ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*, Columbus, Ohio, USA.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proc. of the ACL 2007 Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics, demo sessions*, pages 177–180, Prague, Czech Republic, June.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 1:29, pages 19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Conf. of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, July.

Matthias Paulik, Kay Rottmann, Jan Niehues, Silja Hildebrand, and Stephan Vogel. 2007. The ISL phrase-based mt system for the 2007 ACL workshop on statistical machine translation. In *In Proc. of the ACL 2007 Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver.