NRC's PORTAGE system for WMT 2007

Nicola Ueffing, Michel Simard, Samuel Larkin
Interactive Language Technologies Group
National Research Council Canada
Gatineau, Québec, Canada
firstname.lastname@nrc.gc.ca

Howard Johnson
Interactive Information Group
National Research Council Canada
Ottawa, Ontario, Canada
Howard.Johnson@nrc.gc.ca

Abstract

We present the PORTAGE statistical machine translation system which participated in the shared task of the ACL 2007 Second Workshop on Statistical Machine Translation. The focus of this description is on improvements which were incorporated into the system over the last year. These include adapted language models, phrase table pruning, an IBM1-based decoder feature, and rescoring with posterior probabilities.

1 Introduction

The statistical machine translation (SMT) system PORTAGE was developed at the National Research Council Canada and has recently been made available to Canadian universities and research institutions. It is a state-of-the-art phrase-based SMT system. We will shortly describe its basics in this paper and then highlight the new methods which we incorporated since our participation in the WMT 2006 shared task. These include new scoring methods for phrase pairs, pruning of phrase tables based on significance, a higher-order language model, adapted language models, and several new decoder and rescoring models. PORTAGE was also used in a joint system developed in cooperation with Systran. The interested reader is referred to (Simard et al., 2007).

Throughout this paper, let $s_1^J := s_1 \dots s_J$ denote a source sentence of length J, $t_1^I := t_1 \dots t_I$ a target sentence of length I, and $\tilde{s}$ and $\tilde{t}$ phrases in source and target language, respectively.

2 Baseline

As baseline for our experiments, we used a version of PORTAGE corresponding to its state at the time of the WMT 2006 shared task. We provide a basic description of this system here; for more details see (Johnson et al., 2006).

PORTAGE implements a two-stage translation process: First, the decoder generates N-best lists, using a basic set of models which are then rescored with additional models in a second step. In the baseline system, the decoder uses the following models (or feature functions):

- one or several phrase table(s), which model the translation direction $p(\tilde{s} \mid \tilde{t})$. They are generated from the training corpus via the "diag-and" method (Koehn et al., 2003) and smoothed using Kneser-Ney smoothing (Foster et al., 2006),
- one or several n-gram language model(s) trained with the SRILM toolkit (Stolcke, 2002); in the baseline experiments reported here, we used a trigram model,
- a distortion model which assigns a penalty based on the number of source words which are skipped when generating a new target phrase,
- a word penalty.

These different models are combined log-linearly. Their weights are optimized w.r.t. BLEU score using the algorithm described in (Och, 2003). This is done on the provided development corpus. The search algorithm implemented in the decoder is a dynamic-programming beam-search algorithm.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 185–188,
Prague, June 2007.

After the decoding step, rescoring with additional models is performed. The baseline system generates a 1,000-best list of alternative translations for each source sentence. These lists are rescored with the different models described above, a character penalty, and three different features based on IBM Models 1 and 2 (Brown et al., 1993) calculated in both translation directions. The weights of these additional models and of the decoder models are again optimized to maximize BLEU score.

Note that we did not use the decision-tree-based distortion models described in (Johnson et al., 2006) here because they did not improve translation quality.

In the following subsections, we will describe the new models added to the system for our WMT 2007 submissions.

3 Improvements in PORTAGE

3.1 Phrase translation models

Whereas the phrase tables used in the baseline system contain only one score for each phrase pair, namely conditional probabilities calculated using Kneser-Ney smoothing, our current system combines seven different phrase scores.

First, we used several types of phrase table smoothing in the WMT 2007 system because this proved helpful on other translation tasks: relative frequency estimates, Kneser-Ney- and Zens-Ney-smoothed probabilities (Foster et al., 2006). Furthermore, we added normalized joint probability estimates to the phrase translation model. The other three scores will be explained at the end of this subsection.

We pruned the generated phrase tables following the method introduced in (Johnson et al., 2007). This approach considers all phrase pairs $(\tilde{s}, \tilde{t})$ in the phrase table. The count $C(\tilde{s}, \tilde{t})$ of all sentence pairs containing $(\tilde{s}, \tilde{t})$ is determined, as well as the count of all source/target sentences containing $\tilde{s}/\tilde{t}$. Using these counts, Fisher's exact test is carried out to calculate the significance of the phrase pair. The phrase tables are then pruned based on the p-value. Phrase pairs with low significance, i.e. which are only weakly supported by the training data, are pruned. This reduces the size of the phrase tables to 8-16% on the different language pairs. See (Johnson et al., 2007) for details.

Three additional phrase scores were derived from information on which this pruning is based:

- the significance level (or p-value),
- the number $C(\tilde{s}, \tilde{t})$ of sentence pairs containing the phrase pair, normalized by the number of source sentences containing $\tilde{s}$,
- $C(\tilde{s}, \tilde{t})$, normalized by the number of target sentences containing $\tilde{t}$.

For our submissions, we used the last three phrase scores only when translating the EuroParl data. Initial experiments showed that they do not improve translation quality on the News Commentary data. Apart from this, the systems for both domains are identical.

3.2 Adapted language models

Concerning the language models, we made two changes to our system since WMT 2006. First, we replaced the trigram language model by a 4-gram model trained on the WMT 2007 data. We also investigated the use of a 5-gram, but that did not improve translation quality. Second, we included adapted language models which are specific to the development and test corpora. For each development or test corpus, we built this language model using information retrieval[1] to find relevant sentences in the training data. To this end, we merged the training corpora for EuroParl and News Commentary. The source sentences from the development or test corpus served as individual queries to find relevant training sentence pairs. For each source sentence, we retrieved 10 sentence pairs from the training data and used their target sides as language model training data. On this small corpus, we trained a trigram language model, again using the SRILM toolkit. The feature function weights in the decoder and the rescoring model were optimized using the adapted language model for the development corpus. When translating the test corpus, we kept these weights, but replaced the adapted

[1]We used the lemur toolkit for querying, see http://www.lemurproject.org/

language model by that specific to the test corpus.

3.3 New decoder and rescoring features

We integrated several new decoder and rescoring features into PORTAGE. During decoding, the system now makes use of a feature based on IBM Model 1. This feature calculates the probability of the (partial) translation over the source sentence, using an IBM1 translation model in the direction $p(t_1^I \mid s_1^J)$.

In the rescoring process, we additionally included several types of posterior probabilities. One is the posterior probability of the sentence length over the N-best list for this source sentence. The others are determined on the level of words, phrases, and n-grams, and then combined into a value for the whole sentence. All posterior probabilities are calculated over the N-best list, using the sentence probabilities which the baseline system assigns to the translation hypotheses. For details on the posterior probabilities, see (Ueffing and Ney, 2007; Zens and Ney, 2006). This year, we increased the length of the N-best lists from 1,000 to 5,000.

3.4 Post-processing

For truecasing the translation output, we used the model described in (Agbago et al., 2005). This model uses a combination of statistical components, including an n-gram language model, a case mapping model, and a specialized language model for unknown words. The language model is a 5-gram model trained on the WMT 2007 data. The detokenizer which we used is the one provided for WMT 2007.

4 Experimental results

We submitted results for six of the translation directions of the shared task: French ↔ English, German ↔ English, and Spanish ↔ English.

Table 1 shows the improvements resulting from incorporating new techniques into PORTAGE on the Spanish → English EuroParl task. The baseline system is the one described in section 2. Trained on the 2007 training corpora, this yields a BLEU score of 30.48. Adding the new phrase scores introduced in section 3.1 yields a slight improvement in translation quality. This improvement by itself is not significant, but we observed it consistently across all evaluation metrics and across the different development and test corpora. Increasing the order of the language model and adding an adapted language model specific to the translation input (see section 3.2) improves the BLEU score by 0.6 points. This is the biggest gain we observe from introducing a new method. The incorporation of the IBM1-based decoder feature causes a slight drop in translation quality. This surprised us because we found this feature to be very helpful on the NIST Chinese → English translation task. Adding the posterior probabilities presented in section 3.3 in rescoring and increasing the length of the N-best lists yielded a small, but consistent gain in translation quality. The overall improvement compared to last year's system is around 1 BLEU point. The gain achieved from introducing the new methods by themselves are relatively small, but they add up.

Table 2 shows results on all six language pairs we translated for the shared task. The translation quality achieved on the 2007 test set is similar to that on the 2006 test set. The system clearly performs better on the EuroParl domain than on News Commentary.

Table 2: *Translation quality in terms of BLEU[%] and NIST score on all tasks. Truecased and detokenized translation output.*

		test2006		test2007	
task		BLEU	NIST	BLEU	NIST
Eu	D→E	25.27	6.82	26.02	6.91
	E→D	19.36	5.86	18.94	5.71
	S→E	31.54	7.55	32.09	7.67
	E→S	30.94	7.39	30.92	7.41
	F→E	30.90	7.51	31.90	7.68
	E→F	30.08	7.26	30.06	7.26
NC	D→E	20.23	6.19	23.17	7.10
	E→D	13.84	5.38	16.30	5.95
	S→E	31.07	7.68	31.08	8.11
	E→S	30.79	7.73	32.56	8.25
	F→E	24.97	6.78	26.84	7.47
	E→F	24.91	6.79	26.60	7.24

Table 1: *Effect of integrating new models and methods into the PORTAGE system. Translation quality in terms of BLEU and NIST score, WER and PER on the EuroParl Spanish–English 2006 test set. True-cased and detokenized translation output. Best results printed in boldface.*

system	BLEU[%]	NIST	WER[%]	PER[%]
baseline	30.48	7.44	58.62	42.74
+ new phrase table features	30.66	7.48	58.25	42.46
+ 4-gram LM + adapted LM	31.26	7.53	57.93	42.26
+ IBM1-based decoder feature	31.18	7.51	58.13	42.53
+ refined rescoring	**31.54**	**7.55**	**57.81**	**42.24**

5 Conclusion

We presented the PORTAGE system with which we translated six language pairs in the WMT 2007 shared task. Starting from the state of the system during the WMT 2006 evaluation, we analyzed the contribution of new methods which were incorporated over the last year in detail. Our experiments showed that most of these changes result in (small) improvements in translation quality. In total, we gain about 1 BLEU point compared to last year's system.

6 Acknowledgments

Our thanks go to the PORTAGE team at NRC for their contributions and valuable feedback.

References

A. Agbago, R. Kuhn, and G. Foster. 2005. Truecasing for the Portage system. In *Recent Advances in Natural Language Processing*, pages 21–24, Borovets, Bulgaria, September.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

G. Foster, R. Kuhn, and J. H. Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 53–61, Sydney, Australia, July.

J. H. Johnson, F. Sadat, G. Foster, R. Kuhn, M. Simard, E. Joanis, and S. Larkin. 2006. Portage: with smoothed phrase tables and segment choice models. In *Proc. HLT/NAACL Workshop on Statistical Machine Translation (WMT)*, pages 134–137, New York, NY, June.

H. Johnson, J. Martin, G. Foster, and R. Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing and Conf. on Computational Natural Language Learning (EMNLP-CoNLL)*, to appear, Prague, Czech Republic, June.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

M. Simard, J. Senellart, P. Isabelle, R. Kuhn, J. Stephan, and N. Ueffing. 2007. Knowledge-based translation with statistical phrase-based post-editing. In *Proc. ACL Second Workshop on Statistical Machine Translation (WMT)*, to appear, Prague, Czech Republic, June.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO.

N. Ueffing and H. Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40, March.

R. Zens and H. Ney. 2006. N-gram posterior probabilities for statistical machine translation. In *Proc. HLT/NAACL Workshop on Statistical Machine Translation (WMT)*, pages 72–77, New York, NY, June.