PORTAGE: A Phrase-based Machine Translation System

Fatiha Sadat[+], Howard Johnson[++], Akakpo Agbago[+], George Foster[+], Roland Kuhn[+], Joel Martin[++] and Aaron Tikuisis[*]

[+] NRC Institute for Information Technology
101 St-Jean-Bosco Street
Gatineau, QC K1A 0R6, Canada

[++] NRC Institute for Information Technology
1200 Montreal Road
Ottawa, ON K1A 0R6, Canada

[*] University of Waterloo
200 University Avenue W.,
Waterloo, Ontario, Canada

firstname.lastname@cnrc-nrc.gc.ca aptikuis@uwaterloo.ca

Abstract

This paper describes the participation of the Portage team at NRC Canada in the shared task[1] of ACL 2005 Workshop on Building and Using Parallel Texts. We discuss Portage, a statistical phrase-based machine translation system, and present experimental results on the four language pairs of the shared task. First, we focus on the French-English task using multiple resources and techniques. Then we describe our contribution on the Finnish-English, Spanish-English and German-English language pairs using the provided data for the shared task.

1 Introduction

The rapid growth of the Internet has led to a rapid growth in the need for information exchange among different languages. Machine Translation (MT) and related technologies have become essential to the information flow between speakers of different languages on the Internet. Statistical Machine Translation (SMT), a data-driven approach to producing translation systems, is becoming a practical solution to the longstanding goal of cheap natural language processing.

In this paper, we describe Portage, a statistical phrase-based machine translation system, which we evaluated on all different language pairs that were provided for the shared task. As Portage is a very new system, our main goal in participating in the workshop was to test it out on different language pairs, and to establish baseline performance for the purpose of comparison against other systems and against future improvements. To do this, we used a fairly standard configuration for phrase-based SMT, described in the next section.

Of the language pairs in the shared task, French-English is particularly interesting to us in light of Canada's demographics and policy of official bilingualism. We therefore divided our participation into two parts: one stream for French-English and another for Finnish-, German-, and Spanish-English. For the French-English stream, we tested the use of additional data resources along with hand-coded rules for translating numbers and dates. For the other streams, we used only the provided resources in a purely statistical framework (although we also investigated several automatic methods of coping with Finnish morphology).

The remainder of the paper is organized as follows. Section 2 describes the architecture of the Portage system, including its hand-coded rules for French-English. Experimental results for the four pairs of languages are reported in Section 3. Section 4 concludes and gives pointers to future work.

2 Portage

Portage operates in three main phases: *preprocessing* of raw data into tokens, with translation suggestions for some words or phrases generated by rules; *decoding* to produce one or more translation hypotheses; and error-driven *rescoring* to choose the best final hypothesis. (A fourth *postprocessing* phase was not needed for the shared task.)

[1] http://www.statmt.org/wpt05/mt-shared-task/

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 129–132,
Ann Arbor, June 2005.

2.1 Preprocessing

Preprocessing is a necessary first step in order to convert raw texts in both source and target languages into a format suitable for both model training and decoding (Foster et al., 2003). For the supplied *Europarl* corpora, we relied on the existing segmentation and tokenization, except for French, which we manipulated slightly to bring into line with our existing conventions (e.g., converting *l ' an* into *l' an).* For the *Hansard* corpus used to supplement our French-English resources (described in section 3 below), we used our own alignment based on Moore's algorithm (Moore, 2002), segmentation, and tokenization procedures.

Languages with rich morphology are often problematic for statistical machine translation because the available data lacks instances of all possible forms of a word to efficiently train a translation system. In a language like German, new words can be formed by compounding (writing two or more words together without a space or a hyphen in between). Segmentation is a crucial step in preprocessing languages such as German and Finnish texts.

In addition to these simple operations, we also developed a rule-based component to detect numbers and dates in the source text and identify their translation in the target text. This component was developed on the Hansard corpus, and applied to the French-English texts (i.e. Europarl and Hansard), on the development data in both languages, and on the test data.

2.2 Decoding

Decoding is the central phase in SMT, involving a search for the hypotheses *t* that have highest probabilities of being translations of the current source sentence *s* according to a model for *P(t|s)*. Our model for *P(t|s)* is a log-linear combination of four main components: one or more trigram language models, one or more phrase translation models, a distortion model, and a word-length feature. The trigram language model is implemented in the SRILM toolkit (Stolcke, 2002). The phrase-based translation model is similar to the one described in (Koehn, 2004), and relies on symmetrized IBM model 2 word-alignments for phrase pair induction. The distortion model is also very similar to Koehn's, with the exception of a final cost to account for sentence endings.

To set weights on the components of the log-linear model, we implemented Och's algorithm (Och, 2003). This essentially involves generating, in an iterative process, a set of nbest translation hypotheses that are representative of the entire search space for a given set of source sentences. Once this is accomplished, a variant of Powell's algorithm is used to find weights that optimize BLEU score (Papineni et al, 2002) over these hypotheses, compared to reference translations. Unfortunately, our implementation of this algorithm converged only very slowly to a satisfactory final nbest list, so we used two different ad hoc strategies for setting weights: choosing the best values encountered during the iterations of Och's algorithm (French-English), and a grid search (all other languages).

To perform the actual translation, we used our decoder, *Canoe*, which implements a dynamic-programming beam search algorithm based on that of *Pharaoh* (Koehn, 2004). Canoe is input-output compatible with Pharaoh, with the exception of a few extensions such as the ability to decode either backwards or forwards.

2.3 Rescoring

To improve raw output from Canoe, we used a rescoring strategy: have Canoe generate a list of nbest translations rather than just one, then reorder the list using a model trained with Och's method to optimize BLEU score. This is identical to the final pass of the algorithm described in the previous section, except for the use of a more powerful log-linear model than would have been feasible to use inside the decoder. In addition to the four basic features of the initial model, our rescoring model included IBM2 model probabilities in both directions (i.e., P(s|t) and P(t|s)); and an IBM1-based feature designed to detect whether any words in one language seemed to be left without satisfactory translations in the other language. This *missing-word* feature was also applied in both directions.

3 Experiments on the Shared Task

We conducted experiments and evaluations on Portage using the different language pairs of the shared task. The training data was provided for the shared task as follows:

- Training data of 688,031 sentences in French and English. A similarly sized cor-

pus is provided for Finnish, Spanish and German with matched English translations.
- Development test data of 2,000 sentences in the four languages.

In addition to the provided data, a set of 6,056,014 sentences extracted from Hansard corpus, the official record of Canada's parliamentary debates, was used in both French and English languages. This corpus was used to generate both language and translation models for use in decoding and rescoring.

The development test data was split into two parts: The first part that includes 1,000 sentences in each language with reference translations into English served in the optimization of weights for both the decoding and rescoring models. In this study, number of n-best lists was set to 1,000. The second part, which includes 1,000 sentences in each language with reference translations into English, was used in the evaluation of the performance of the translation models.

3.1 Experiments on the French-English Task

Our goal for this language pair was to conduct experiments on Portage for a comparative study exploiting and combining different resources and techniques:

1. Method E is based on the *Europarl* corpus as training data,
2. Method E-H is based on both *Europarl* and *Hansard* corpora as training data,
3. Method E-p is based on the *Europarl* corpus as training data and *parsing numbers and dates* in the preprocessing phase,
4. Method E-H-p is based on both *Europarl* and *Hansard* corpora as training data and *parsing numbers and date* in the preprocessing phase.

Results are shown in Table 1 for the French-English task. The first column of Table 1 indicates the method, the second column gives results for decoding with Canoe only, and the third column for decoding and rescoring with Canoe. For comparison between the four methods, there was an improvement in terms of BLEU scores when using two language models and two translation models generated from Europarl and Hansard corpora; however, parsing numbers and dates had a negative impact on the translation models. The best BLEU score for our participation at the French-English task was 29.53.

Method	Decoding	Decoding+Rescoring
E	27.71	29.22
E-H	28.71	**29.53**
E-p	26.45	28.21
E-H-p	28.29	28.56

Table 1. BLEU scores for the French-English test sentences

A noteworthy feature of these results is that the improvement given by the out-of-domain Hansard corpus was very slight. Although we suspect that somewhat better performance could have been achieved by better weight optimization, this result clearly underscores the importance of matching training and test domains. A related point is that our number and date translation rules actually caused a performance drop due to the fact that they were optimized for typographical conventions prevalent in Hansard, which are quite different from those used in Europarl.

Our best result ranked third in the shared WPT05 French-English task , with a difference of 0.74 in terms of BLEU score from the first ranked participant, and a difference of 0.67 in terms of BLEU score from the second ranked participant.

3.2 Experiments on other Pairs of Languages

The WPT05 workshop provides a good opportunity to achieve our benchmarking goals with corpora that provide challenging difficulties. German and Finnish are languages that make considerable use of compounding. Finnish, in addition, has a particularly complex morphology that is organized on principles that are quite different from any in English. This results in much longer word forms each of which occurs very infrequently.

Our original intent was to propose a number of possible statistical approaches to analyzing and splitting these word forms and improving our results. Since none of these yielded results as good as the baseline, we will continue this work until we understand what is really needed. We also care very much about translating between French and English in Canada and plan to spend a lot of extra effort on difficulties that occur in this case. Translation between Spanish and English is also becoming more important as a result of increased trade within North America but also functions as a good counterpoint for French-English.

Language Pair	Decoding+Rescoring
Finnish-English	20.95
German-English	23.21
Spanish English	29.08

Table 2 BLEU scores for the Finnish-English, German-English and Spanish-English test sentences

To establish our baseline, the only preprocessing we did was lowercasing (using the provided tokenization). Canoe was run without any special settings, although weights for distortion, word penalty, language model, and translation model were optimized using a grid search, as described above. Rescoring was also done, and usually resulted in at least an extra BLEU point.

Our final results are shown in Table 2. Ranks at the shared WPT05 Finnish-, German-, and Spanish-English tasks were assigned as second, third and fourth, with differences of 1.06, 1.87 and 1.56 in terms of BLEU scores, respectively, compared to the first ranked participant.

4 Conclusion

We have reported on our participation in the shared task of the ACL 2005 Workshop on Building and Using Parallel Texts, conducting evaluations of Portage, our statistical machine translation system, on all four language pairs. Our best BLEU scores for the French-, Finnish-, German-, and Spanish-English at this stage were 29.5, 20.95, 23.21 and 29.08, respectively. In total, eleven teams took part at the shared task and most of them submitted results for all pairs of languages. Our results distinguished the NRC team at the third, second, third and fourth ranks with slight differences with the first ranked participants.

A major goal of this work was to evaluate Portage at its first stage of implementation on different pairs of languages. This evaluation has served to identify some problems with our system in the areas of weight optimization and number and date rules. It has also indicated the limits of using out-of-domain corpora, and the difficulty of morphologically complex languages like Finnish.

Current and planned future work includes the exploitation of comparable corpora for statistical machine translation, greater use of morphological knowledge, and better features for nbest rescoring.

References

Andreas Stolcke. 2002. *SRILM - an Extensible Language Modeling Toolkit*. In ICSLP-2002, 901-904.

Franz Josef Och, Hermann Ney. 2000. *Improved Statistical Alignment Models*. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, October 2000, 440-447.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, Dragomir Radev. 2004. *A Smorgasbord of Features for Statistical Machine Translation*. In Proceeding of the HLT/NAACL 2004, Boston, MA, May 2004.

George Foster, Simona Gandrabur, Philippe Langlais, Pierre Plamondon, Graham Russell and Michel Simard. 2003. *Statistical Machine Translation: Rapid Development with Limited Resources.* In Proceedings of MT Summit IX 2003, New Orleans, September.

Kevin Knight, Ishwar Chander, Matthew Haines, Vasileios Hatzivassiloglou, Eduard Hovy, Masayo Iida, Steve K. Luk, Richard Whitney, and Kenji Yamada. 1995. *Filling Knowledge Gaps in a Broad-Coverage MT System*. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 1995.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics ACL, Philadelphia, July 2002, pp. 311-318.

Moore, Robert. 2002. *Fast and Accurate Sentence Alignment of Bilingual Corpora.* In Machine Translation: From Research to Real Users (Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, pp. 135-244.

Och, F. J. and H. Ney. 2002. *Discriminative Training and Maximum Entropy Models for Statistical Machine Translation*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp. 295–302.

Franz Josef Och, 2003. *Minimum Error Rate Training for Statistical Machine Translation*. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, July.

Philipp Koehn. 2002. *Europarl: A multilingual corpusfor evaluation of machine translation*. Ms., University of Southern California.

Philipp Koehn. 2004. *Pharaoh:* a *Beam Search Decoder for Phrase-based Statistical Machine Translation Models.* In Proceedings of the Association for Machine Translation in the Americas AMTA 2004.