PORTAGE: with Smoothed Phrase Tables and Segment Choice Models

Howard Johnson
National Research Council
Institute for Information Technology
Interactive Information
1200 Montreal Road
Ottawa, ON, Canada K1A 0R6
Howard.Johnson@cnrc-nrc.gc.ca

Fatiha Sadat, George Foster, Roland Kuhn,
Michel Simard, Eric Joanis and Samuel Larkin
National Research Council
Institute for Information Technology
Interactive Language Technologies
101 St-Jean-Bosco Street
Gatineau, QC, Canada K1A 0R6
firstname.lastname@cnrc-nrc.gc.ca

Abstract

Improvements to Portage and its participation in the shared task of NAACL 2006 Workshop on Statistical Machine Translation are described. Promising ideas in phrase table smoothing and global distortion using feature-rich models are discussed as well as numerous improvements in the software base.

1 Introduction

The statistical machine translation system Portage is participating in the NAACL 2006 Workshop on Statistical Machine Translation. This is a good opportunity to do benchmarking against a publicly available data set and explore the benefits of a number of recently added features.

Section 2 describes the changes that have been made to Portage in the past year that affect the participation in the 2006 shared task. Section 3 outlines the methods employed for this task and extensions of it. In Section 4 the results are summarized in tabular form. Following these, there is a conclusions section that highlights what can be gleaned of value from these results.

2 Portage

Because this is the second participation of Portage in such a shared task, a description of the base system can be found elsewhere (Sadat et al, 2005). Briefly, Portage is a research vehicle and development prototype system exploiting the state-of-the-art in statistical machine translation (SMT). It uses a custom built decoder followed by a rescoring module that adjusts weights based on a number of features defined on the source sentence. We will devote space to discussing changes made since the 2005 shared task.

2.1 Phrase-Table Smoothing

Phrase-based SMT relies on conditional distributions $p(s|t)$ and $p(t|s)$ that are derived from the joint frequencies $c(s,t)$ of source/target phrase pairs observed in an aligned parallel corpus. Traditionally, relative-frequency estimation is used to derive conditional distributions, ie $p(s|t) = c(s,t)/\sum_s c(s,t)$. However, relative-frequency estimation has the well-known problem of favouring rare events. For instance, any phrase pair whose constituents occur only once in the corpus will be assigned a probability of 1, almost certainly higher than the probabilities of pairs for which much more evidence exists. During translation, rare pairs can directly compete with overlapping frequent pairs, so overestimating their probabilities can significantly degrade performance.

To address this problem, we implemented two simple smoothing strategies. The first is based on the Good-Turing technique as described in (Church and Gale, 1991). This replaces each observed joint frequency c with $c_g = (c+1)n_{c+1}/n_c$, where n_c is the number of distinct pairs with frequency c (smoothed for large c). It also assigns a total count mass of n_1 to unseen pairs, which we distributed in proportion to the frequency of each conditioning

Proceedings of the Workshop on Statistical Machine Translation, pages 134–137,
New York City, June 2006. ©2006 Association for Computational Linguistics

phrase. The resulting estimates are:

$$p_g(s|t) = \frac{c_g(s,t)}{\sum_s c_g(s,t) + p(t)n_1},$$

where $p(t) = c(t)/\sum_t c(t)$. The estimates for $p_g(t|s)$ are analogous.

The second strategy is Kneser-Ney smoothing (Kneser and Ney, 1995), using the interpolated variant described in (Chen and Goodman., 1998):[1]

$$p_k(s|t) = \frac{c(s,t) - D + D\, n_{1+}(*,t)\, p_k(s)}{\sum_s c(s,t)}$$

where $D = n_1/(n_1 + 2n_2)$, $n_{1+}(*,t)$ is the number of distinct phrases s with which t co-occurs, and $p_k(s) = n_{1+}(s,*)/\sum_s n_{1+}(s,*)$, with $n_{1+}(s,*)$ analogous to $n_{1+}(*,t)$.

Our approach to phrase-table smoothing contrasts to previous work (Zens and Ney, 2004) in which smoothed phrase probabilities are constructed from word-pair probabilities and combined in a log-linear model with an unsmoothed phrase-table. We believe the two approaches are complementary, so a combination of both would be worth exploring in future work.

2.2 Feature-Rich DT-based distortion

In a recent paper (Kuhn et al, 2006), we presented a new class of probabilistic ”Segment Choice Models” (SCMs) for distortion in phrase-based systems. In some situations, SCMs will assign a better distortion score to a drastic reordering of the source sentence than to no reordering; in this, SCMs differ from the conventional penalty-based distortion, which always favours less rather than more distortion.

We developed a particular kind of SCM based on decision trees (DTs) containing both questions of a positional type (e.g., questions about the distance of a given phrase from the beginning of the source sentence or from the previously translated phrase) and word-based questions (e.g., questions about the presence or absence of given words in a specified phrase).

The DTs are grown on a corpus consisting of segment-aligned bilingual sentence pairs. This segment-aligned corpus is obtained by training a phrase translation model on a large bilingual corpus and then using it (in conjunction with a distortion penalty) to carry out alignments between the phrases in the source-language sentence and those in the corresponding target-language sentence in a second bilingual corpus. Typically, the first corpus (on which the phrase translation model is trained) is the same as the second corpus (on which alignment is carried out). To avoid overfitting, the alignment algorithm is leave-one-out: statistics derived from a particular sentence pair are not used to align that sentence pair.

Note that the experiments reported in (Kuhn et al, 2006) focused on translation of Chinese into English. The interest of the experiments reported here on WMT data was to see if the feature-rich DT-based distortion model could be useful for MT between other language pairs.

3 Application to the Shared Task: Methods

3.1 Restricted Resource Exercise

The first exercise that was done is to replicate the conditions of 2005 as closely as possible to see the effects of one year of research and development. The second exercise was to replicate all three of these translation exercises using the 2006 language model, and to do the three exercises of translating out of English into French, Spanish, and German. This was our baseline for other studies. A third exercise involved modifying the generation of the phrase-table to incorporate our Good-Turing smoothing. All six language pairs were re-processed with these phrase-tables. The improvement in the results on the devtest set were compelling. This became the baseline for further work. A fourth exercise involved replacing penalty-based distortion modelling with the feature-rich decision-tree based distortion modelling described above. A fifth exercise involved the use of a Kneser-Ney phrase-table smoothing algorithm as an alternative to Good-Turing.

For all of these exercises, 1-best results after decoding were calculated as well as rescoring on 1000-best lists of results using 12 feature functions (13 in the case of decision-tree based distortion modelling). The results submitted for the shared task

[1] As for Good-Turing smoothing, this formula applies only to pairs s, t for which $c(s,t) > 0$, since these are the only ones considered by the decoder.

were the results of the third and fourth exercises where rescoring had been applied.

3.2 Open Resource Exercise

Our goal in this exercise was to conduct a comparative study using additional training data for the French-English shared task. Results of WPT 2005 showed an improvement of at least 0.3 BLEU point when exploiting different resources for the French-English pair of languages. In addition to the training resources used in WPT 2005 for the French-English task, i.e. Europarl and Hansard, we used a bilingual dictionary, *Le Grand Dictionnaire Terminologique* (GDT) [2] to train translation models and the English side of the UN parallel corpus (LDC2004E13) to train an English language model. Integrating terminological lexicons into a statistical machine translation engine is not a straightforward operation, since we cannot expect them to come with attached probabilities. The approach we took consists on viewing all translation candidates of each source term or phrase as equiprobable (Sadat et al, 2006).

In total, the data used in this second part of our contribution to WMT 2006 is described as follows: (1) A set of 688,031 sentences in French and English extracted from the *Europarl parallel corpus* (2) A set of 6,056,014 sentences in French and English extracted from the *Hansard parallel corpus*, the official record of Canada's parliamentary debates. (3) A set of 701,709 sentences in French and English extracted from the bilingual dictionary *GDT*. (4) Language models were trained on the French and English parts of the Europarl and Hansard. We used the provided Europarl corpus while omitting data from Q4/2000 (October-December), since it is reserved for development and test data. (5) An additional English language model was trained on 128 million words of the *UN Parallel corpus*.

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, aujourd ' hui into aujourd'hui).

For the Hansard corpus used to supplement our French-English resources, we used our own alignment based on Moore's algorithm, segmentation, and tokenization procedures. English preprocessing simply included lower-casing, separating punctuation from words and splitting off 's.

4 Results

The results are shown in Table 1. The numbers shown are BLEU scores. The MC rows correspond to the multi-corpora results described in the open resource exercise section above. All other rows are from the restricted resource exercise.

The devtest results are the scores computed before the shared-task submission and were used to drive the choice of direction of the research. The test results were computed after the shared-task submission and serve for validation of the conclusions.

We believe that our use of multiple training corpora as well as our re-tokenization for French and an enhanced language model resulted in our overall success in the English-French translation track. The results for the in-domain test data puts our group at the top of the ranking table drawn by the organizers (first on Adequacy and fluency and third on BLEU scores).

5 Conclusion

Benchmarking with same language model and parameters as WPT05 reproduces the results with a tiny improvement. The larger language model used in 2006 for English yields about half a BLEU. Good-Turing phrase table smoothing yields roughly half a BLEU point. Kneser-Ney phrase table smoothing yields between a third and half a BLEU point more than Good-Turing. Decision tree based distortion yields a small improvement for the devtest set when rescoring was not used but failed to show improvement on the test set.

In summary, the results from phrase-table smoothing are extremely encouraging. On the other hand, the feature-rich decision tree distortion modelling requires additional work before it provides a good pay-back. Fortunately we have some encouraging avenues under investigation. Clearly there is more work needed for both of these areas.

Acknowledgements

We wish to thank Aaron Tikuisis and Denis Yuen for important contributions to the Portage code base

[2] http://www.granddictionnaire.com/

Table 1: Restricted and open resource results

	fr ⟶ en	es ⟶ en	de ⟶ en	en ⟶ fr	en ⟶ es	en ⟶ de
devtest: with rescoring						
WPT05	29.32	29.08	23.21			
LM-2005	29.30	29.21	23.41			
LM-2006	29.88	29.54	23.94	30.43	28.81	17.33
GT-PTS	30.35	29.84	24.60	30.89	29.54	17.62
GT-PTS+DT-dist	30.09	29.44	24.62	31.06	29.46	**17.84**
KN-PTS	**30.55**	**30.12**	**24.66**	**31.28**	**29.90**	17.78
MC WPT05	29.63					
MC	30.09			31.30		
MC+GT-PTS	**30.75**			**31.37**		
devtest: 1-best after decoding						
LM-2006	28.59	28.45	23.22	29.22	28.30	16.94
GT-PTS	29.23	28.91	**23.67**	30.07	28.86	17.32
GT-PTS+DT-dist	29.48	29.07	23.50	30.22	29.46	17.42
KN-PTS	**29.77**	**29.76**	23.27	**30.73**	**29.62**	**17.78**
MC WPT05	28.71					
MC	29.63			31.01		
MC+GT-PTS	**29.90**			**31.22**		
test: with rescoring						
LM-2006	26.64	28.43	21.33	28.06	28.01	15.19
GT-PTS	27.19	28.95	21.91	28.60	28.83	15.38
GT-PTS+DT-dist	26.84	28.56	21.84	28.56	28.59	15.45
KN-PTS	**27.40**	**29.07**	**21.98**	**28.96**	**29.06**	**15.64**
MC	26.95			29.12		
MC+GT-PTS	**27.10**			**29.46**		
test: 1-best after decoding						
LM-2006	25.35	27.25	20.46	27.20	27.18	14.60
GT-PTS	25.95	28.07	21.06	27.85	27.96	15.05
GT-PTS+DT-dist	25.86	28.04	20.74	27.85	27.97	14.92
KN-PTS	**26.83**	**28.66**	**21.36**	**28.62**	**28.71**	**15.42**
MC	26.70			28.74		
MC+GT-PTS	**26.81**			**29.03**		

and the OQLF (Office Québécois de la Langue Française) for permission to use the GDT.

References

S. F. Chen and J. T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

K. Church and W. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer speech and language*, 5(1):19–54.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1995*, pages 181–184, Detroit, Michigan. IEEE.

R. Kuhn, D. Yuen, M. Simard, G. Foster, P. Paul, E. Joanis and J. H. Johnson. 2006. Segment Choice Models: Feature-Rich Models for Global Distortion in Statistical Machine Translation (accepted for publication in HLT-NAACL conference, to be held June 2006).

F. Sadat, J. H. Johnson, A. Agbago, G. Foster, R. Kuhn, J. Martin and A. Tikuisis. 2005. PORTAGE: A Phrase-based Machine Translation System In *Proc. ACL 2005 Workshop on building and using parallel texts.* Ann Arbor, Michigan.

F. Sadat, G. Foster and R. Kuhn. 2006. Système de traduction automatique statistique combinant différentes ressources. In *Proc. TALN 2006 (Traitement Automatique des Langues Naturelles).* Leuven, Belgium, April 10-13, 2006.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. Human Language Technology Conference / North American Chapter of the ACL*, Boston, May.