European Language Translation with Weighted Finite State Transducers: The CUED MT System for the 2008 ACL Workshop on SMT

Graeme Blackwood, Adrià de Gispert, Jamie Brunning, William Byrne
Machine Intelligence Laboratory
Department of Engineering, Cambridge University
Trumpington Street, Cambridge, CB2 1PZ, U.K.
{gwb24|ad465|jjjb2|wjb31}@cam.ac.uk

Abstract

We describe the Cambridge University Engineering Department phrase-based statistical machine translation system for Spanish-English and French-English translation in the ACL 2008 Third Workshop on Statistical Machine Translation Shared Task. The CUED system follows a generative model of translation and is implemented by composition of component models realised as Weighted Finite State Transducers, without the use of a special-purpose decoder. Details of system tuning for both Europarl and News translation tasks are provided.

1 Introduction

The Cambridge University Engineering Department statistical machine translation system follows the Transducer Translation Model (Kumar and Byrne, 2005; Kumar et al., 2006), a phrase-based generative model of translation that applies a series of transformations specified by conditional probability distributions and encoded as Weighted Finite State Transducers (Mohri et al., 2002).

The main advantages of this approach are its modularity, which facilitates the development and evaluation of each component individually, and its implementation simplicity which allows us to focus on modeling issues rather than complex decoding and search algorithms. In addition, no special-purpose decoder is required since standard WFST operations can be used to obtain the 1-best translation or a lattice of alternative hypotheses. Finally, the system architecture readily extends to speech translation, in which input ASR lattices can be translated in the same way as for text (Mathias and Byrne, 2006).

This paper reviews the first participation of CUED in the ACL Workshop on Statistical Machine Translation in 2008. It is organised as follows. Firstly, section 2 describes the system architecture and its main components. Section 3 gives details of the development work conducted for this shared task and results are reported and discussed in section 4. Finally, in section 5 we summarise our participation in the task and outline directions for future work.

2 The Transducer Translation Model

Under the Transducer Translation Model, the generation of a target language sentence t_1^J starts with the generation of a source language sentence s_1^I by the source language model $P_G(s_1^I)$. Next, the source language sentence is segmented into phrases according to the unweighted uniform phrasal segmentation model $P_W(u_1^K, K|s_1^I)$. This source phrase sequence generates a reordered target language phrase sequence according to the phrase translation and reordering model $P_R(x_1^K|u_1^K)$. Next, target language phrases are inserted into this sequence according to the insertion model $P_\Phi(v_1^R|x_1^K, u_1^K)$. Finally, the sequence of reordered and inserted target language phrases are transformed to word sequences t_1^J under the target phrasal segmentation model $P_\Omega(t_1^J|v_1^R)$. These component distributions together form a joint distribution over the source and target language sentences and their possible intermediate phrase sequences as $P(t_1^J, v_1^R, x_1^K, u_1^K, s_1^I)$.

In translation under the generative model, we start with the target sentence t_1^J in the foreign language

Proceedings of the Third Workshop on Statistical Machine Translation, pages 131–134,
Columbus, June 2008.

and search for the best source sentence $\hat{s}_1^I$. Encoding each distribution as a WFST leads to a model of translation as the series of compositions

$$L = G \circ W \circ R \circ \Phi \circ \Omega \circ T \tag{1}$$

in which T is an acceptor for the target language sentence and L is the word lattice of translations obtained during decoding. The most likely translation $\hat{s}_1^I$ is the path in L with least cost.

2.1 TTM Reordering Model

The TTM reordering model associates a jump sequence with each phrase pair. For the experiments described in this paper, the jump sequence is restricted such that only adjacent phrases can be swapped; this is the MJ1 reordering model of (Kumar and Byrne, 2005). Although the reordering probability for each pair of phrases could be estimated from word-aligned parallel data, we here assume a uniform reordering probability p tuned as described in section 3.1. Figure 1 shows how the MJ1 reordering model for a pair of phrases **x1** and **x2** is implemented as a WFST.

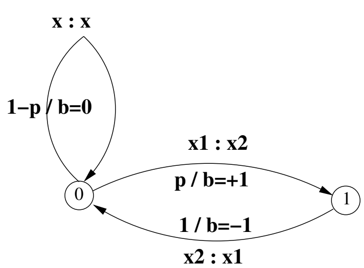

Figure 1: *The uniform MJ1 reordering transducer.*

3 System Development

CUED participated in two of the WMT shared task tracks: French→English and Spanish→English. For both tracks, primary and contrast systems were submitted. The primary submission was restricted to only the parallel and language model data distributed for the shared task. The contrast submission incorporates large additional quantities of English monolingual training text for building the second-pass language model described in section 3.2.

Table 1 summarises the parallel training data, including the total number of sentences, total number of words, and lower-cased vocabulary size. The Spanish and French parallel texts each contain approximately 5% News Commentary data; the rest is Europarl data. Various single-reference development and test sets were provided for each of the tracks. However, the 2008 evaluation included a new News task, for which no corresponding development set was available.

	sentences	words	vocab
FR	1.33M	39.9M	124k
EN		36.4M	106k
ES	1.30M	38.2M	140k
EN		35.7M	106k

Table 1: *Parallel corpora statistics.*

All of the training and system tuning was performed using lower-cased data. Word alignments were generated using GIZA++ (Och and Ney, 2003) over a stemmed version of the parallel text. Stems for each language were obtained using the Snowball stemmer[1]. After unioning the Viterbi alignments, the stems were replaced with their original words, and phrase-pairs of up to five foreign words in length were extracted in the usual fashion (Koehn et al., 2003).

3.1 System Tuning

Minimum error training (Och, 2003) under BLEU (Papineni et al., 2001) was used to optimise the feature weights of the decoder with respect to the *dev2006* development set. The following features are optimized:

- Language model scale factor
- Word and phrase insertion penalties
- Reordering scale factor
- Insertion scale factor
- Translation model scale factor: u-to-v
- Translation model scale factor: v-to-u
- Three phrase pair count features

The phrase-pair count features track whether each phrase-pair occurred once, twice, or more than twice

[1]Available at http://snowball.tartarus.org

in the parallel text (Bender et al., 2007). All decoding and minimum error training operations are performed with WFSTs and implemented using the OpenFST libraries (Allauzen et al., 2007).

3.2 English Language Models

Separate language models are used when translating the Europarl and News sets. The models are estimated using SRILM (Stolcke, 2002) and converted to WFSTs for use in TTM translation. We use the offline approximation in which failure transitions are replaced with epsilons (Allauzen et al., 2003).

The Europarl language model is a Kneser-Ney (Kneser and Ney, 1995) smoothed default-cutoff 5-gram back-off language model estimated over the concatenation of the Europarl and News language model training data. The News language model is created by optimising the interpolation weights of two component models with respect to the News Commentary development sets since we believe these more closely match the *newstest2008* domain. The optimised interpolation weights were 0.44 for the Europarl corpus and 0.56 for the much smaller News Commentary corpus. For our contrast submission, we rescore the first-pass translation lattices with a large zero-cutoff stupid-backoff (Brants et al., 2007) language model estimated over approximately five billion words of newswire text.

4 Results and Discussion

Table 2 reports lower-cased BLEU scores for the French→English and Spanish→English Europarl and News translation tasks. The NIST scores are also provided in parentheses. The row labelled "TTM+MET" shows results obtained after TTM translation and minimum error training, i.e. our primary submission constrained to use only the data distributed for the task. The row labelled "+5gram" shows translation results obtained after rescoring with the large zero-cutoff 5-gram language model described in section 3.2. Since this includes additional language model data, it represents the CUED contrast submission.

Translation quality for the ES→EN task is slightly higher than that of FR→EN. For Europarl translation, most of the additional English language model training data incorporated into the 5-gram rescoring step is out-of-domain and so does not substantially improve the scores. Rescoring yields an average gain of just +0.5 BLEU points.

Translation quality is significantly lower in both language pairs for the new *news2008* set. Two factors may account for this. The first is the change in domain and the fact that no training or development set was available for the News translation task. Secondly, the use of a much freer translation in the single News reference, which makes it difficult to obtain a good BLEU score. However, the second-pass 5-gram language model rescoring gains are larger than those observed in the Europarl sets, with approximately +1.7 BLEU points for each language pair. The additional in-domain newswire data clearly helps to improve translation quality.

Finally, we use a simple 3-gram casing model trained on the true-case workshop distributed language model data, and apply the SRILM `disambig` tool to restore true-case for our final submissions. With respect to the lower-cased scores, true-casing drops around 1.0 BLEU in the Europarl task, and around 1.7 BLEU in the News Commentary and News tasks.

5 Summary

We have reviewed the Cambridge University Engineering Department first participation in the workshop on machine translation using a phrase-based SMT system implemented with a simple WFST architecture. Results are largely competitive with the state-of-the-art in this task.

Future work will examine whether further improvements can be obtained by incorporating additional features into MET, such as the word-to-word Model 1 scores or phrasal segmentation models. The MJ1 reordering model could also be extended to allow for longer-span phrase movement. Minimum Bayes Risk decoding, which has been applied successfully in other tasks, could also be included.

The difference in the gains from 5-gram lattice rescoring suggests that, particularly for Europarl translation, it is important to ensure the language model data is in-domain. Some form of count mixing or alternative language model adaptation techniques may prove useful for unconstrained Europarl translation.

Task		dev2006	devtest2006	test2007	**test2008**	**newstest2008**
FR→EN	TTM+MET	31.92 (7.650)	32.51 (7.719)	32.94 (7.805)	32.83 (7.799)	19.58 (6.108)
	+5gram	32.51 (7.744)	32.96 (7.797)	33.33 (7.880)	33.03 (7.856)	21.22 (6.311)
ES→EN	TTM+MET	33.11 (7.799)	32.25 (7.649)	32.90 (7.766)	33.11 (7.859)	20.99 (6.308)
	+5gram	33.30 (7.835)	32.96 (7.740)	33.55 (7.857)	33.47 (7.893)	22.83 (6.513)

Table 2: *Translation results for the Europarl and News tasks for various dev sets and the 2008 test sets.*

Acknowledgements

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

References

Cyril Allauzen, Mehryar Mohri, and Brian Roark. 2003. Generalized algorithms for constructing statistical language models. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 557–564.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFST: a general and efficient weighted finite-state transducer library. In *Proceedings of the 9th International Conference on Implementation and Application of Automata*, pages 11–23. Springer.

Oliver Bender, Evgeny Matusov, Stefan Hahn, Sasa Hasan, Shahram Khadivi, and Hermann Ney. 2007. The RWTH Arabic-to-English spoken language translation system. In *Proceedings of the 2007 Automatic Speech Understanding Workshop*, pages 396–401.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 858–867.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing*, pages 181–184.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA.

Shankar Kumar and William Byrne. 2005. Local phrase reordering models for statistical machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 161–168.

Shankar Kumar, Yonggang Deng, and William Byrne. 2006. A weighted finite state transducer translation template model for statistical machine translation. *Natural Language Engineering*, 12(1):35–75.

Lambert Mathias and William Byrne. 2006. Statistical phrase-based speech translation. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. In *Computer Speech and Language*, volume 16, pages 69–88.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*.