The CUED HiFST System for the WMT10 Translation Shared Task

Juan Pino Gonzalo Iglesias[‡1] **Adrià de Gispert**
Graeme Blackwood Jamie Brunning William Byrne
Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K.
{jmp84,gi212,ad465,gwb24,jjjb2,wjb31}@eng.cam.ac.uk
[‡] Department of Signal Processing and Communications, University of Vigo, Vigo, Spain

Abstract

This paper describes the Cambridge University Engineering Department submission to the Fifth Workshop on Statistical Machine Translation. We report results for the French-English and Spanish-English shared translation tasks in both directions. The CUED system is based on HiFST, a hierarchical phrase-based decoder implemented using weighted finite-state transducers. In the French-English task, we investigate the use of context-dependent alignment models. We also show that lattice minimum Bayes-risk decoding is an effective framework for multi-source translation, leading to large gains in BLEU score.

1 Introduction

This paper describes the Cambridge University Engineering Department (CUED) system submission to the ACL 2010 Fifth Workshop on Statistical Machine Translation (WMT10). Our translation system is HiFST (Iglesias et al., 2009a), a hierarchical phrase-based decoder that generates translation lattices directly. Decoding is guided by a CYK parser based on a synchronous context-free grammar induced from automatic word alignments (Chiang, 2007). The decoder is implemented with Weighted Finite State Transducers (WFSTs) using standard operations available in the OpenFst libraries (Allauzen et al., 2007). The use of WFSTs allows fast and efficient exploration of a vast translation search space, avoiding search errors in decoding. It also allows better integration with other steps in our translation pipeline such as 5-gram language model (LM) rescoring and lattice minimum Bayes-risk (LMBR) decoding.

[1]Now a member of the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K.

	# Sentences	# Tokens	# Types
(A)Europarl+News-Commentary			
FR	1.7 M	52.4M	139.7k
EN		47.6M	121.6k
(B)Europarl+News-Commentary+UN			
FR	8.7 M	277.9M	421.0k
EN		241.4M	482.1k
(C)Europarl+News-Commentary+UN+Giga			
FR	30.2 M	962.4M	2.4M
EN		815.3M	2.7M

Table 1: Parallel data sets used for French-to-English experiments.

We participated in the French-English and Spanish-English translation shared tasks in each translation direction. This paper describes the development of these systems. Additionally, we report multi-source translation experiments that lead to very large gains in BLEU score.

The paper is organised as follows. Section 2 describes each step in the development of our system for submission, from pre-processing to post-processing. Section 3 presents and discusses results and Section 4 describes an additional experiment on multi-source translation.

2 System Development

We built three French-English and two Spanish-English systems, trained on different portions of the parallel data sets available for this shared task. Statistics for the different parallel sets are summarised in Tables 1 and 2. No additional parallel data was used. As will be shown, the largest parallel corpus gave the best results in French, but this was not the case in Spanish.

2.1 Pre-processing

The data was minimally cleaned by replacing HTML-related metatags by their corresponding

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 161–166,
Uppsala, Sweden, 11-16 July 2010.

	# Sentences	# Tokens	# Types
(A) Europarl + News-Commentary			
SP	1.7M	49.4M	167.2k
EN		47.0M	122.7k
(B) Europarl + News-Commentary + UN			
SP	6.5M	205.6M	420.8k
EN		192.0M	402.8k

Table 2: Parallel data sets used for Spanish-to-English experiments.

UTF8 token (e.g., replacing "&amp" by "&") as this interacts with tokenization. Data was then tokenized and lowercased, so mixed case is added as post-processing.

2.2 Alignments

Parallel data was aligned using the MTTK toolkit (Deng and Byrne, 2005). In the English-to-French and English-to-Spanish directions, we trained a word-to-phrase HMM model with maximum phrase length of 2. In the French to English and Spanish to English directions, we trained a word-to-phrase HMM Model with a bigram translation table and maximum phrase length of 4.

We also trained context-dependent alignment models (Brunning et al., 2009) for the French-English medium-size (B) dataset. The context of a word is based on its part-of-speech and the part-of-speech tags of the surrounding words. These tags were obtained by applying the TnT Tagger (Brants, 2000) for English and the TreeTagger (Schmid, 1994) for French. Decision tree clustering based on optimisation of the EM auxiliary function was used to group contexts that translate similarly. Unfortunately, time constraints prevented us from training context-dependent models for the larger (C) dataset.

2.3 Language Model

For each target language, we used the SRILM Toolkit (Stolcke, 2002) to estimate separate 4-gram LMs with Kneser-Ney smoothing (Kneser and Ney, 1995), for each of the corpora listed in Tables 3, 4 and 5. The LM vocabulary was adjusted to the parallel data set used. The component models of each language pair were then interpolated to form a single LM for use in first-pass translation decoding. For French-to-English translation, the interpolation weights were optimised for perplexity on a development set.

Corpus	# Sentences	# Tokens
EU + NC + UN	9.0M	246.4M
CNA	1.3M	34.8M
LTW	12.9M	298.7M
XIN	16.0M	352.5M
AFP	30.4M	710.6M
APW	62.1M	1268.6M
NYT	73.6M	1622.5M
Giga	21.4M	573.8M
News	48.7M	1128.4M
Total	275.4M	6236.4M

Table 3: English monolingual training corpora.

Corpus	# Sentences	# Tokens
EU + NC + UN	9.0M	282.8
AFP	25.2M	696.0M
APW	12.7M	300.6M
News	15.2M	373.5M
Giga	21.4M	684.4M
Total	83.5 M	2337.3M

Table 4: French monolingual training corpora.

Corpus	# Sentences	# Tokens
NC + News	4.0M	110.8M
EU + Gigaword (5g)	249.4M	1351.5M
Total	253.4 M	1462.3M

Table 5: Spanish monolingual training corpora.

The Spanish-English first pass LM was trained directly on the NC+News portion of monolingual data, as we did not find improvements by using Europarl. The second pass rescoring LM used all available data.

2.4 Grammar Extraction and Decoding

After unioning the Viterbi alignments, phrase-based rules of up to five source words in length were extracted, hierarchical rules with up to two non-contiguous non-terminals in the source side were then extracted applying the restrictions described in (Chiang, 2007). For Spanish-English and French-English tasks, we used a shallow-1 grammar where hierarchical rules are allowed to be applied only once on top of phrase-based rules. This has been shown to perform as well as a fully hierarchical grammar for a Europarl Spanish-English task (Iglesias et al., 2009b).

For translation, we used the HiFST de-

coder (Iglesias et al., 2009a). HiFST is a hierarchical decoder that builds target word lattices guided by a probabilistic synchronous context-free grammar. Assuming $\mathbf{N}$ to be the set of non-terminals and $\mathbf{T}$ the set of terminals or words, then we can define the grammar as a set $\mathbf{R} = \{R^r\}$ of rules $R^r : N \rightarrow \langle \gamma^r, \alpha^r \rangle$ / p^r, where $N \in \mathbf{N}$; and $\gamma, \alpha \in \{\mathbf{N} \cup \mathbf{T}\}^+$.

HiFST translates in three steps. The first step is a variant of the CYK algorithm (Chappelier and Rajman, 1998), in which we apply hypothesis recombination without pruning. Only the source language sentence is parsed using the corresponding source-side context-free grammar with rules $N \rightarrow \gamma$. Each cell in the CYK grid is specified by a non-terminal symbol and position: (N, x, y), spanning s_x^{x+y-1} on the source sentence $s_1 ... s_J$.

For the second step, we use a recursive algorithm to construct word lattices with all possible translations produced by the hierarchical rules. Construction proceeds by traversing the CYK grid along the back-pointers established in parsing. In each cell (N, x, y) of the CYK grid, we build a target language word lattice $\mathcal{L}(N, x, y)$ containing every translation of s_x^{x+y-1} from every derivation headed by N. For efficiency, this lattice can use pointers to lattices on other cells of the grid.

In the third step, we apply the word-based LM via standard WFST composition with failure transitions, and perform likelihood-based pruning (Allauzen et al., 2007) based on the combined translation and LM scores.

As explained before, we are using shallow-1 hierarchical grammars (de Gispert et al., 2010) in our experiments for WMT2010. One very interesting aspect is that HiFST is able to build exact search spaces with this model, i.e. there is no pruning in search that may lead to spurious undergeneration errors.

2.5 Parameter Optimisation

Minimum error rate training (MERT) (Och, 2003) under the BLEU score (Papineni et al., 2001) optimises the weights of the following decoder features with respect to the *newstest2008* development set: target LM, number of usages of the glue rule, word and rule insertion penalties, word deletion scale factor, source-to-target and target-to-source translation models, source-to-target and target-to-source lexical models, and three binary rule count features inspired by Bender et al. (2007) indicating whether a rule occurs once, twice, or more than twice in the parallel training data.

2.6 Lattice Rescoring

One of the advantages of HiFST is direct generation of large translation lattices encoding many alternative translation hypotheses. These first-pass lattices are rescored with second-pass higher-order LMs prior to LMBR.

2.6.1 5-gram LM Lattice Rescoring

We build sentence-specific, zero-cutoff stupid-backoff (Brants et al., 2007) 5-gram LMs estimated over approximately 6.2 billion words for English, 2.3 billion words for French, and 1.4 billion words for Spanish. For the English-French task, the second-pass LM training data is the same monolingual data used for the first-pass LMs (as summarised in Tables 3, 4). The Spanish second-pass 5-gram LM includes an additional 1.4 billion words of monolingual data from the Spanish GigaWord Second Edition (Mendonca et al., 2009) and Europarl, which were not included in the first-pass LM (see Table 5).

2.6.2 LMBR Decoding

Minimum Bayes-risk (MBR) decoding (Kumar and Byrne, 2004) over the full evidence space of the 5-gram rescored lattices was applied to select the translation hypothesis that maximises the conditional expected gain under the linearised sentence-level BLEU score (Tromble et al., 2008; Blackwood and Byrne, 2010). The unigram precision p and average recall ratio r were set as described in Tromble et al. (2008) using the *newstest2008* development set.

2.7 Hypothesis Combination

Linearised lattice minimum Bayes-risk decoding (Tromble et al., 2008) can also be used as an effective framework for multiple lattice combination (de Gispert et al., 2010). For the French-English language pair, we used LMBR to combine translation lattices produced by systems trained on alternative data sets.

2.8 Post-processing

For both Spanish-English and French-English systems, the recasing procedure was performed with the SRILM toolkit. For the Spanish-English system, we created models from the GigaWord set corresponding to each system output language.

Task	Configuration	*newstest2008*	*newstest2009*	*newstest2010*
FR → EN	HiFST (A)	23.4	26.4	–
	HiFST (B)	24.0	27.3	–
	HiFST $(B)^{CD}$	24.2	27.6	28.0
	+5g+LMBR	**24.6**	**28.4**	**28.9**
	HiFST (C)	24.7	28.4	28.5
	+5g+LMBR	**25.3**	**29.1**	**29.3**
	LMBR $(B)^{CD}$+(C)	**25.6**	**29.3**	**29.6**
EN → FR	HiFST (A)	22.5	24.2	–
	HiFST (B)	23.4	24.8	–
	HiFST $(B)^{CD}$	23.3	24.8	26.7
	+5g+LMBR	**23.7**	**25.3**	**27.1**
	HiFST (C)	23.6	25.6	27.4
	+5g+LMBR	**23.9**	**25.8**	**27.8**
	LMBR $(B)^{CD}$+(C)	**24.2**	**26.1**	**28.2**

Table 6: Translation Results for the French-English (FR-EN) language pair, shown in single-reference lowercase IBM BLEU. Bold results correspond to submitted systems.

For the French-English system, the English model was trained using the monolingual News corpus and the target side of the News-Commentary corpus, whereas the French model was trained using all available constrained French data.

English, Spanish and French outputs were also detokenized before submission. In French, words separated by apostrophes were joined.

3 Results and Discussion

French–English Language Pair

Results are reported in Table 6. We can see that using more parallel data consistently improves performance. In the French-to-English direction, the system HiFST (B) improves over HiFST (A) by +0.9 BLEU and HiFST (C) improves over HiFST (B) by +1.1 BLEU on the *newstest2009* development set prior to any rescoring. The same trend can be observed in the English-to-French direction (+0.6 BLEU and +0.8 BLEU improvement). The use of context dependent alignment models gives a small improvement in the French-to-English direction: system $(B)^{CD}$ improves by +0.3 BLEU over system (B) on *newstest2009*. In the English-to-French direction, there is no improvement nor degradation in performance. 5-gram and LMBR rescoring also give consistent improvement throughout the datasets. Finally, combination between the medium-size system $(B)^{CD}$ and the full-size system (C) gives further small gains in BLEU over LMBR on each individual system.

Spanish–English Language Pair

Results are reported in Table 7. We report experimental results on two systems. The HiFST(A) system is built on the Europarl + News-Commentary training set. Systems HiFST (B),(B2) and (B3) use UN data in different ways. System (B) simply uses all the data for the standard rule extraction procedure. System HiFST (B2) includes UN data to build alignment models and therefore reinforce alignments obtained from smaller dataset (A), but extracts rules only from dataset (A). HiFST (B3) combines hierarchical phrases extracted for system (A) with phrases extracted from system (B). Unfortunately, these three larger data strategies lead to degradation over using only the smaller dataset (A). For this reason, our best systems only use the Euparl + News-Commentary parallel data. This is surprising given that additional data was helpful for the French-English task. Solving this issue is left for future work.

4 Multi-Source Translation Experiments

Multi-source translation (Och and Ney, 2001; Schroeder et al., 2009) is possible whenever multiple translations of the source language input sentence are available. The motivation for multi-source translation is that some of the ambiguity that must be resolved in translating between one pair of languages may not be present in a different pair. In the following experiments, multiple LMBR is applied for the first time to the task of multi-source translation.

Task	Configuration	*newstest2008*	*newstest2009*	*newstest2010*
SP → EN	HiFST (A)	24.6	26.0	29.1
	+5g+LMBR	**25.4**	**27.0**	**30.5**
	HiFST (B)	23.7	25.4	–
	HiFST (B2)	24.3	25.7	–
	HiFST (B3)	24.2	25.6	–
EN → SP	HiFST (A)	23.9	24.5	28.0
	+5g+LMBR	**24.7**	**25.5**	**29.1**

Table 7: Translation Results for the Spanish-English (SP-EN) language pair, shown in lowercase IBM BLEU. Bold results correspond to submitted systems.

Configuration		*newstest2008*	*newstest2009*	*newstest2010*
FR→EN	HiFST+5g	24.8	28.5	28.8
	+LMBR	25.3	29.0	29.2
ES→EN	HiFST+5g	25.2	26.8	30.1
	+LMBR	25.4	26.9	30.3
FR→EN + ES→EN	LMBR	27.2	30.4	32.0

Table 8: Lowercase IBM BLEU for single-system LMBR and multiple LMBR multi-source translation of French (FR) and Spanish (ES) into English (EN).

Separate second-pass 5-gram rescored lattices $\mathcal{E}_{\text{FR}}$ and $\mathcal{E}_{\text{ES}}$ are generated for each test set sentence using the French-to-English and Spanish-to-English HiFST translation systems. The MBR hypothesis space is formed as the union of these lattices. In a similar manner to MBR decoding over multiple k-best lists in de Gispert et al. (2009), the path posterior probability of each n-gram u required for linearised LMBR is computed as a linear interpolation of the posterior probabilities according to each individual lattice so that $p(u|\mathcal{E}) = \lambda_{\text{FR}}\, p(u|\mathcal{E}_{\text{FR}}) + \lambda_{\text{ES}}\, p(u|\mathcal{E}_{\text{ES}})$, where $p(u|\mathcal{E})$ is the sum of the posterior probabilities of all paths containing the n-gram u. The interpolation weights $\lambda_{\text{FR}} + \lambda_{\text{ES}} = 1$ are optimised for BLEU score on the development set *newstest2008*.

The results of single-system and multi-source LMBR decoding are shown in Table 8. The optimised interpolation weights were $\lambda_{\text{FR}} = 0.55$ and $\lambda_{\text{ES}} = 0.45$. Single-system LMBR gives relatively small gains on these test sets. Much larger gains are obtained through multi-source MBR combination. Compared to the best of the single-system 5-gram rescored lattices, the BLEU score improves by +2.0 for *newstest2008*, +1.9 for *newstest2009*, and +1.9 for *newstest2010*. For scoring with respect to a single reference, these are very large gains indeed.

5 Summary

We have described the CUED submission to WMT10 using HiFST, a hierarchical phrase-based translation system. Results are very competitive in terms of automatic metric for both English-French and English-Spanish tasks in both directions. In the French-English task, we have seen that the UN and Giga additional parallel data are helpful. It is surprising that UN data did not help for the Spanish-English language pair.

Future work includes investigating this issue, developing detokenization tailored to each output language and applying context dependent alignment models to larger parallel datasets.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7-ICT-2009-4) under grant agreement number 247762, and was supported in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. Gonzalo Iglesias was supported by the Spanish Government research grant BES-2007-15956 (projects TEC2006-13694-C03-03 and TEC2009-14094-C04-04).

References

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 CIAA*, pages 11–23.

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 ASRU*, pages 396–401.

Graeme Blackwood and William Byrne. 2010. Efficient Path Counting Transducers for Minimum Bayes-Risk Decoding of Statistical Machine Translation Lattices (to appear). In *Proceedings of the ACL*.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of EMNLP-ACL*, pages 858–867.

Thorsten Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proceedings of ANLP*, pages 224–231, April.

Jamie Brunning, Adrià de Gispert, and William Byrne. 2009. Context-dependent alignment models for statistical machine translation. In *Proceedings of HLT/NAACL*, pages 110–118.

Jean-Cédric Chappelier and Martin Rajman. 1998. A generalized CYK algorithm for parsing stochastic CFG. In *Proceedings of TAPD*, pages 133–137.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions. In *Proceedings of HLT/NAACL, Companion Volume: Short Papers*, pages 73–76.

Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, and William Byrne. 2010. Hierarchical phrase-based translation with weighted finite state transducers and shallow-n grammars (to appear). In *Computational Linguistics*.

Yonggang Deng and William Byrne. 2005. HMM Word and Phrase Alignment for Statistical Machine Translation. In *Proceedings of HLT/EMNLP*, pages 169–176.

Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, and William Byrne. 2009a. Hierarchical phrase-based translation with weighted finite state transducers. In *Proceedings of NAACL*, pages 433–441.

Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, and William Byrne. 2009b. The HiFST System for the EuroParl Spanish-to-English Task. In *Proceedings of SEPLN*, pages 207–214.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of ICASSP*, volume 1, pages 181–184.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 169–176.

Angelo Mendonca, David Graff, and Denise DiPersio. 2009. Spanish Gigaword Second Edition, Linguistic Data Consortium.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Machine Translation Summit 2001*, pages 253–258.

Franz J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word Lattices for Multi-Source Translation. In *Proceedings of EACL*, pages 719–727.

Andreas Stolcke. 2002. SRILM–An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP*, volume 3, pages 901–904.

Roy W. Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of EMNLP*, pages 620–629.