

# CUED Submission for the WMT10 Translation Shared Task

Juan Pino<sup>‡</sup> Gonzalo Iglesias\* Adrià de Gispert<sup>‡</sup> Graeme Blackwood<sup>‡</sup> Jamie Brunning<sup>‡</sup> William Byrne<sup>‡</sup>

\* University of Vigo. Dept. of Signal Processing and Communications. Vigo, Spain

{giglesia}@gts.tsc.uvigo.es

<sup>‡</sup> University of Cambridge. Dept. of Engineering. CB2 1PZ Cambridge, U.K.

{jmp84,gwb24,ad465,jjyb2,wjb31}@eng.cam.ac.uk

## Abstract

This paper describes the Cambridge University Engineering Department (CUED) system for the ACL 2010 fifth workshop on statistical machine translation (WMT10). We participated in the French-English and Spanish-English translation shared tasks in both directions. The CUED system is a hierarchical phrase-based system that uses finite-state transducers and lattice rescoring. In the French-English task, we investigated the use of context-dependent alignments.

## 1 Introduction

This paper describes the CUED system submission for WMT10. We use HiFST (Iglesias et al., 2009a), a hierarchical phrase-based system that builds word translation lattices guided by a CYK parser based on a synchronous context-free grammar induced from automatic word alignments. The decoder is implemented with Weighted Finite State Transducers (WFSTs) using standards operations available in the OpenFst libraries (Allauzen et al., 2007). (TODO: OK TO COPY ?) The use of WFSTs allows better integration with other steps in the system pipeline such as 5-gram and lattice minimum Bayes-risk decoding.

We participated in the French-English and Spanish-English translation shared tasks in both directions. We present results for systems build on part of the available data and all available data. The French-English system investigates the use of context-dependent alignments (Brunnering et al., 2009).

The paper is organized 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

### 2.1 Resources

We built two French-English systems. The first one used all parallel data available. The second one used only part of the data. Statistics for the two subsets are summarized in Table 2 and 1 (TODO: Gonzalo, could you give me stats for Spanish ?). The first subset will be designated by *medium-size*, the second subset by *full-size* in the remainder of the paper.

Europarl + News-commentary + UN + Giga			
	# sentences	# tokens	# types
FR	30,159,797	962,412,088	2,363,335
EN	30,159,797	815,280,736	2,743,933

Table 1: Parallel Corpus Statistics for all data

Europarl + News-commentary + UN			
	# sentences	# tokens	# types
FR	8,789,381	277,974,129	421,039
EN	8,789,381	241,443,499	482,137
SP			
EN			

Table 2: Parallel Corpus Statistics for a subset of the data

### 2.2 Pre-processing

The data was cleaned by replacing certain character strings with other strings, for example replacing “&amp” by “&”. The French side of the parallel data was tokenized using the script provided by the organizers. The script was modified to take into account French abbreviations. The English side was tokenized using a tokenizer borrowed from RWTH Aachen University (TODO: how to cite this ?). Both sides were lowercased and words mapped to integers.

## 2.3 Alignments

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

We also trained context-dependent alignments (Brunning et al., 2009) for the medium size dataset. The context of a word is based on the part-of-speech of the surrounding words. For English, we used the TNT Tagger (Brants, 2000). For French, we used TreeTagger (Schmid, 1994).

## 2.4 Language Model

We built language models for each of the components in Tables 3 and 4 using the SRILM toolkit (Stolcke, 2002). Each component is a Kneser-Ney (Kneser and Ney, 1995) smoothed default-cutoff 4-gram back-off language model. We then interpolated the different components and optimized them on a development set. The medium-size vocabulary language model was tuned on the `news-test2008` development set. The full-size vocabulary language model was tuned on the `newssyscomb2009` development set. (TODO: Spanish)

	# sentences	# tokens
AFP	30363051	710619911
APW	62144946	1268636668
CNA	1312836	34809056
LTW	12929402	298733005
NYT	73652616	1622508489
XIN	15981967	352505566
E + N-C + UN	9010154	246433375
News	48653884	1128362488
Giga	21370416	573825496
Total	275419272	6236434054

Table 3: Monolingual Data Statistics for English

## 2.5 Grammar Extraction

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 de-

	# sentences	# tokens
AFP	25239870	696009044
APW	12694860	300579763
Euparl	1855589	54533270
News-commentary	97033	2721535
UN	7041240	225614976
News	15234997	373454705
Giga	21370416	684375365
Total	83534005	2337288658

Table 4: Monolingual Data Statistics for French

scribed in (Chiang, 2007). (TODO: all restriction (even boundary) applied ?) We used a shallow grammar where hierarchical rules are allowed to be applied only once on top of phrase-based rules. Iglesias and colleagues (Iglesias et al., 2009b) showed that for a Europarl Spanish-English task, using a full hiero grammar did not improve performance over using a shallow grammar.

## 2.6 Decoding

For decoding, we used our new system called HiFST (Iglesias et al., 2009a). In brief, HiFST is a hierarchical decoder that builds target word lattices guided by a synchronous context-free grammar consisting of a set  $\mathbf{R} = \{R^r\}$  of rules  $R^r : N \rightarrow \langle \gamma^r, \alpha^r \rangle / p^r$ . A priori,  $N$  represents any non-terminal; in this paper,  $N$  can be any non-terminal corresponding to a shallow-1 hierarchical grammar (de Gispert et al., 2010). As usual, the special glue rules  $S \rightarrow \langle X, X \rangle$  and  $S \rightarrow \langle S X, S X \rangle$  are included.  $\mathbf{T}$  denotes the terminals (words), and the grammar builds parses based on strings  $\gamma, \alpha \in \{\mathbf{N} \cup \mathbf{T}\}^+$ , where we follow Chiang’s general restrictions to the grammar (2007).

The system performs translation in three main steps. The first step is a variant of the classic Cocke-Younger-Kasami (CYK) algorithm closely related to CYK+ (Chappelier and Rajman, 1998), for which hypothesis recombination without pruning is performed and back-pointers are maintained. Although the model is a synchronous grammar, in this stage only the source language sentence is parsed using the corresponding context-free grammar with rules  $N \rightarrow \gamma$ . Each cell in the CYK grid is specified by a non-terminal symbol and position in the CYK grid:  $(N, x, y)$ , which spans  $s_x^{x+y-1}$  on the source sentence  $s_1 \dots s_J$ .

For the second step, we use a recursive algo-

rithm 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  $\mathcal{L}(N, x, y)$ . This lattice contains every translation of  $s_x^{x+y-1}$  from every derivation headed by  $N$ . For each rule in this cell, a simple acceptor is built based on the target side, as the result of standard concatenation of acceptors that encode either terminals or non-terminals. A word is encoded trivially with an acceptor of two states binded by a single transition arc. In turn, a non-terminal corresponds to a lattice retrieved by means of its back-pointer to a low-level cell. If this low-level lattice has not been required previously, it has to be built first. Once we have all the acceptors (one per rule) that apply to  $(N, x, y)$ , we obtain the cell lattice  $\mathcal{L}(N, x, y)$  by unioning all these acceptors. For a source sentence with  $J$  words, the lattice we are looking for is at the top cell  $(S, 1, J)$ .

The final translation lattice  $\mathcal{L}(S, 1, J)$  can become very large after the pointer arcs are expanded. Therefore, in the third step we apply a word-based language model via WFST composition, and perform likelihood-based pruning (Alauzen et al., 2007) based on the combined translation and language model scores.

This method can be seen as a generalization of the  $k$ -best algorithm with cube pruning (Chiang, 2007), as the performance of this cube pruning search is clearly limited by the size  $k$  of each  $k$ -best list. For small tasks where  $k$  is sufficiently large compared to the number of translations of each derivation, search could be exhaustive. On the other hand, for reasonably large tasks, the inventory of hierarchical phrases is much bigger than standard phrase pair tables, and using a large  $k$  is impossible without exponentially increasing memory requirements and decoding time. In practice, values of (no more than)  $k = 1000$  or  $k = 10000$  are used. This results in search errors. Search errors have two negative consequences. Clearly, translation quality is undermined as the obtained first-best hypothesis is suboptimal given the models. Additionally, the quality of the obtained  $k$ -best list is also suboptimal, which limits the margin of gain potentially achieved by subsequent re-ranking strategies (such as high order language model rescoring or Minimum Bayes Risk).

## 2.7 Tuning

We use Minimum Error Rate Training (Och, 2003) under the BLEU to optimize the following features on the `news-test2008` development set:

- target language model
- 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
- three binary features for the number of times a rule appears in training (one, two or more than 2)

## 2.8 Rescoring

### 2.8.1 5-gram rescoring

We build sentence-specific zero-cutoff stupid-backoff (Brants et al., 2007) 5-gram language models using all data described in Tables 3 and 4. (TODO: cite paper, which one ?)

### 2.8.2 MBR rescoring

5-gram rescoring outputs a lattice that is rescored using lattice minimum Bayes' risk (MBR) decoding (Tromble et al., 2008).

## 2.9 Combination

In the French-English task, two systems were trained. The first system was trained on the medium size dataset with context-dependent alignments. The second system was trained on the full dataset with context independent alignments. Both system produce a lattice after 5-gram rescoring. These two lattices are combined using a variant of lattice MBR decoding.

## 3 Results and Discussion

Results are presented in Table 5. Translation output is uppercased and detokenized and scored using lower case IBM BLEU. The English-French detokenizer is the same as the Spanish-English detokenizer except that it joins words separated by apostrophes. This can give important gains as apostrophes usually follow function words in French. We can see that using more parallel data

improves performance after rescoring by +0.7 BLEU in the French-to-English direction for *newstest2009* and by +0.5 BLEU in the English-to-French direction for *news-test2008*. We note that 5g-rescoring gains are relatively small in the French-to-English direction. This might be due to having only one reference and because the first pass language model is already strong. In the English-to-French direction, 5g-rescoring is not helpful because it was applied before detokenization, therefore the word insertion penalty might not be appropriate. LMBR rescoring on the other hand gives +0.5 BLEU improvement for *newstest2009* in both French-to-English and English-to-French directions. Finally, combination between the medium-size system and the full-size system gives further small gains in BLEU. Unfortunately, the size of the data did not allow us to test the effect of context-dependent alignments on the full dataset.

For Spanish-to-English we report experiments that also use rules extracted from UN data. The first experiment (HiFST+UN) consists of simply using all this data for the general extraction procedure. For the second one (HiFST+UN.v2) we just used the UN data to reinforce best alignments obtained from Euparl+News (Adria?). The third strategy combines hierarchical phrases extracted for HiFST.UN with phrases extracted from HiFST+UN.v2. Unfortunately, any of these three strategies leads to a degradation in performance. For this reason, our best systems only use the training data from Euparl+News-Commentary. For both Spanish/English systems, the recasing procedure is performed with the SRILM toolkit. For this we create models from the Gigaword set corresponding to the respective target language.

#### 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. Linearised lattice minimum Bayes-risk decoding (Tromble et al., 2008) can be used as an effective framework for multiple lattice combination (de Gispert et al., 2010); in the following experiments, multiple lattice MBR is applied for the

first time to the task of multi-source translation.

Separate second-pass 5-gram rescored lattices  $\mathcal{E}_{FR}$  and  $\mathcal{E}_{ES}$  are generated for each test set sentence using the French→English and Spanish→English HiFST translation systems. The MBR hypothesis space is formed as the union of these two lattices  $\mathcal{E}_{FR} \oplus \mathcal{E}_{ES}$ . The path posterior probability of each  $n$ -gram  $u$  required for linearised lattice MBR is computed as a linear interpolation of the probabilities according to each individual lattice so that

$$p(u|\mathcal{E}) = \lambda_{FR} p(u|\mathcal{E}_{FR}) + \lambda_{ES} p(u|\mathcal{E}_{ES}), \quad (1)$$

where  $p(u|\mathcal{E})$  is the sum of the posterior probabilities of all paths containing the  $n$ -gram  $u$ . The interpolation weights  $\lambda_{FR} + \lambda_{ES} = 1$  are optimised for BLEU score on the development set *newstest2008*.

The results of single-system and multi-source lattice MBR decoding are shown in Table 6. The optimised interpolation weights were  $\lambda_{FR} = 0.55$  and  $\lambda_{ES} = 0.45$ . 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 *newstest-2008*, +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 French-English tasks in both directions.

Future work includes proper use of detokenization for French in MERT and applying context dependent alignment models to larger parallel datasets.

#### Acknowledgments

#### 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.
- Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language

Task		news-test2008	newstest2009	newstest2010
FR → EN	HiFST-medium+MERT	24.2	27.6	28.0
	+5g	24.2	27.9	28.6
	+lbr	24.6	28.4	28.9
	HiFST-full+MERT	24.7	28.4	28.5
	+5g	24.9	28.6	29.0
	+lbr	25.3	29.1	29.3
	combination	25.6	29.3	29.6
EN → FR	HiFST-medium+MERT	23.3	24.8	26.7
	+5g	23.1	24.8	26.6
	+lbr	23.7	25.3	27.1
	HiFST-full+MERT	23.6	25.6	27.4
	+5g	23.4	25.3	27.2
	+lbr	23.9	25.8	27.8
	combination	24.2	26.1	28.2
SP → EN	HiFST +UN	23.7	25.4	-
	HiFST +UN.v2	24.3	25.7	-
	HiFST +UN.v3	24.2	25.6	-
	HiFST noUN	24.6	26.0	-
	+5g	25.2	26.8	-
	+5g+lbr	24.4	<b>27.0</b>	-
EN → SP	HiFST +noUN	23.9	24.5	-
	+5g	24.4	25.1	-
	+5g+lbr	24.7	<b>25.5</b>	-

Table 5: Results (TODO: include newstest2010 ?)

models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867.

Thorsten Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 224–231, Seattle, Washington, USA, April. Association for Computational Linguistics.

Jamie Brunning, Adrià de Gispert, and William Byrne. 2009. Context-dependent alignment models for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 110–118, Boulder, Colorado, June. Association for Computational Linguistics.

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, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Barga, and William Byrne. 2010.

Hierarchical phrase-based translation with weighted finite state transducers and shallow-n grammars (accepted for publication April 2010). In *Computational Linguistics*. Association for Computational Linguistics.

Yonggang Deng and William Byrne. 2005. Hmm word and phrase alignment for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 169–176, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Barga, 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. Barga, 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 *Acoustics, Speech, and Signal Processing*, volume 1.

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

- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.
- 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 the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 719–727, Athens, Greece, March. Association for Computational Linguistics.
- Andreas Stolcke. 2002. Srilm—an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 3.
- Roy W. Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice minimum bayes-risk decoding for statistical machine translation. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Morristown, NJ, USA. Association for Computational Linguistics.

Configuration		newstest2008		newstest2009		newstest2010		
		BLEU	NIST	BLEU	NIST	BLEU	NIST	
FR→EN	HiFST+5g	24.8	6.87	28.5	7.50	28.8	7.59	
	+LMBR	25.3	6.93	29.0	7.58	29.2	7.64	
ES→EN	HiFST+5g	25.2	6.88	26.8	7.38	30.1	7.78	
	+LMBR	25.4	6.89	26.9	7.40	30.3	7.80	
FR→EN + ES→EN		LMBR	27.2	7.17	30.4	7.78	32.0	8.00

Table 6: BLEU and NIST scores for single-system lattice MBR and multiple lattice minimum Bayes-risk multi-source translation of French (FR) and Spanish (ES) into English (EN).