The CUED HiFST System for WMT10

J. Pino, G. Iglesias, A. de Gispert, G. W. Blackwood, J. Brunning, W. J. Byrne

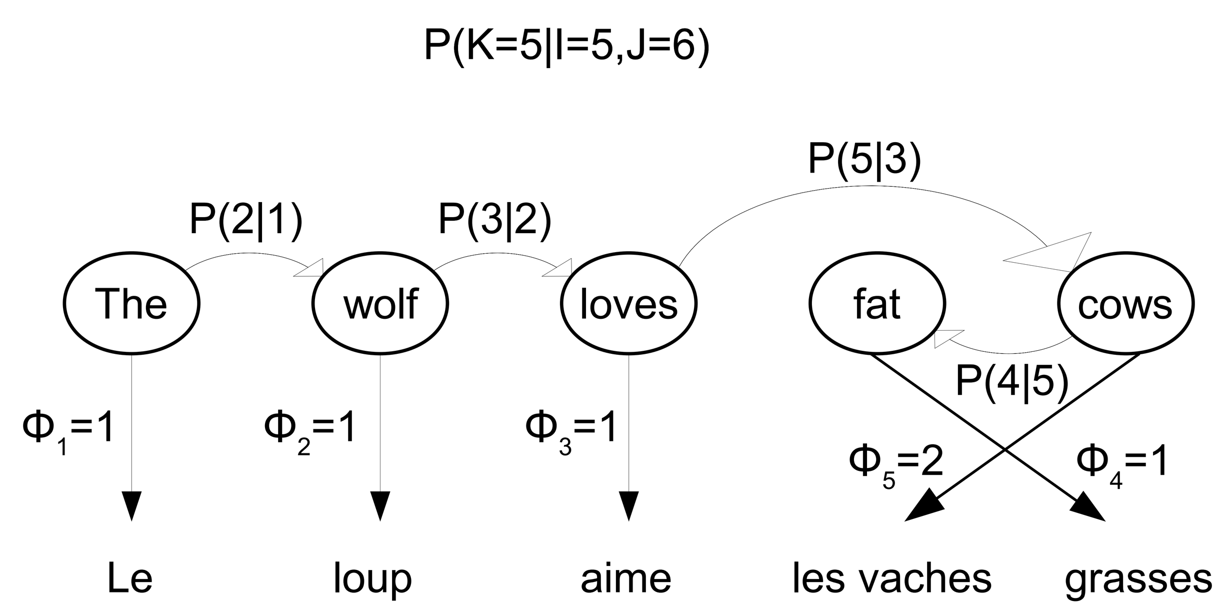
Department of Engineering, University of Cambridge

SUMMARY

- Cambridge University Engineering Department submission to WMT10.
- Translation tasks: FR↔EN and SP↔EN shared translation tasks.
- Decoder: HiFST (Iglesias et al., NAACL'09), a hierarchical phrase-based decoder implemented using WFSTs.
- Shallow hierarchical grammar that requires no pruning in search.
- Investigate the use of context-dependent alignment models in the FR-EN system.
- Additional experiment on multi-source translation: lattice minimum Bayes-risk decoding is an effective framework for multi-source translation, leading to large gains in BLEU score.

TRANSLATION PIPELINE

- **Preprocessing**
  HTML tags, tokenize, lowercase.
- **Alignment**: MTTK (Deng and Byrne, EMNLP'05)
  Word-to-phrase HMM

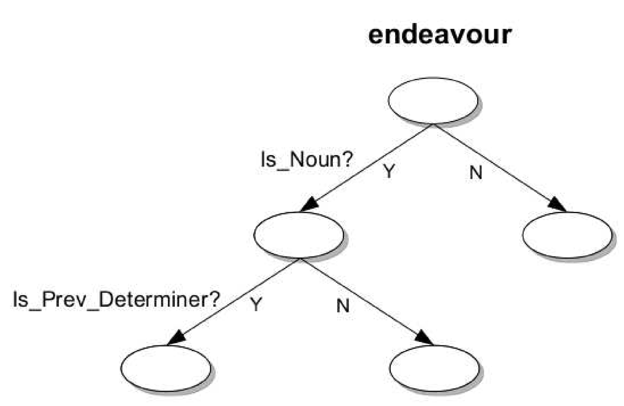


Context dependent HMM (Brunning et al., NAACL'09)

$$p(f_1^J, a_1^J | e_1^I, c_1^J) = \prod_{j=1}^{J} a(a_j | a_{j-1}, I) t(f_j | e_{a_j}, \mu(c_{a_j}))$$

c_{a_j}: context defined as part-of-speech tags of e_{a_j}, e_{a_j-1}, e_{a_j+1}

μ: clustering of context with decision trees



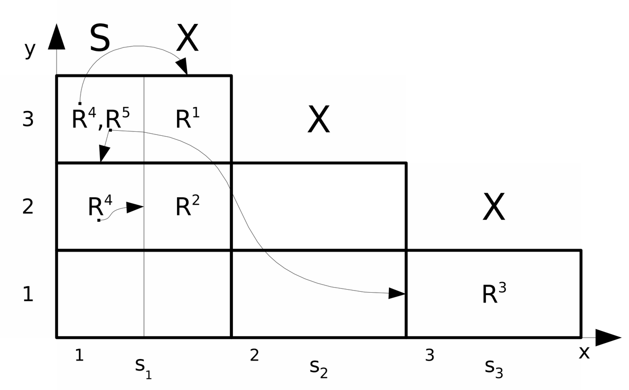
- **1st Pass Language Model**: 4-gram with Kneser-Ney smoothing, in FR-EN interpolate 9 components and optimize weights on a dev set.
- **Grammar Extraction**
  - *Viterbi union*
  - *Rule extraction*: standard phrase and rule extraction, rule filtering by pattern (Iglesias et al., EACL'09).
  - *Shallow-1 grammar* (Iglesias et al., EACL'09):

    Hierarchical rule non terminals must be rewritten with phrase-based rules (degree of rule nesting is 1). Previously shown to work well for the SP-EN language pair.

$$S \rightarrow \langle X, X \rangle$$
$$S \rightarrow \langle S\ X, S\ X \rangle$$
$$X \rightarrow \langle \gamma_s, \alpha_s \rangle (\gamma_s, \alpha_s \in (T \cup \{V\})^+)$$
$$X \rightarrow \langle V, V \rangle$$
$$V \rightarrow \langle s, t \rangle (s \in T^+, t \in T^*)$$

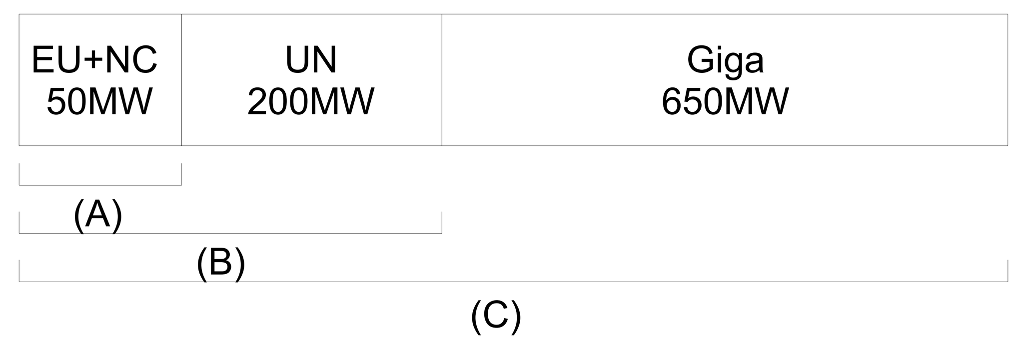
- **Hierarchical Lattice-Based Decoding**
  - Build target language lattice by traversing the CYK grid built on the source.
  - Compose target language translation lattice with language model acceptor.
  - No pruning in search, full exploration of the search space: lattice + shallow-1 grammar.

$$R^1 : X \rightarrow \langle s_1\ s_2\ s_3, t_1\ t_2 \rangle$$
$$R^2 : X \rightarrow \langle s_1\ s_2, t_7\ t_8 \rangle$$
$$R^3 : X \rightarrow \langle s_3, t_9 \rangle$$
$$R^4 : S \rightarrow \langle X, X \rangle$$
$$R^5 : S \rightarrow \langle S\ X, S\ X \rangle$$



- **Features (MERT)**: target LM, number of usage of the glue rule, word and rule insertion penalty, word deletion penalty, source-to-target and target-to-source translation models, source-to-target and target-to-source lexical models, rule count features.
- **5g-Rescoring**: build stupid-backoff 5-gram language models (Brants et al., EMNLP'07) on all available monolingual data.
- **Lattice MBR (LMBR) Rescoring**: maximize conditional expected gain under linearised sentence-level BLEU score (Tromble et al., EMNLP'08; Blackwood et al., ACL'10). Also used for hypothesis combination.
- **Postprocessing**: uppercase, detokenize, join apostrophes in French.

EXPERIMENTS AND RESULTS



- CD: use context dependent alignment models
- (B2): alignment with (B), rules extracted from (A)
- (B3): flat phrases extracted from (B), hierarchical phrases extracted from (A)

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**
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**

- More parallel data helps for FR-EN
- Context dependent models help in the FR→EN direction
- Rescoring methods give consistent gains

MULTI-SOURCE TRANSLATION

- LMBR decoding (de Gispert et al., CL'10) in FR-EN union SP-EN lattices
- Linear interpolation:

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

- Very large gains with respect to best individual systems in combination

Task	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**

CONCLUSIONS

- Very competitive results
- Additional parallel data helps for FR-EN case but not for SP-EN
- LMBR shown to be a very effective framework for multi-source translation