The ITC-irst Statistical Machine Translation System for IWSLT-2004

N. Bertoldi, R. Cattoni, M. Cettolo, M. Federico

ITC-irst

Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), Italy

{bertoldi,cattoni,cettolo,federico}@itc.it

Outline

- The ITC-irst SMT System
  - Log-linear Model
  - Phrase-based Model
  - Decoding
  - System Architecture
- Experiments for IWSLT-2004
  - Selection of Training Data
  - Chinese Segmentation
  - Official Results

Log-linear model for SMT

Maximum Entropy framework for word-alignment MT approach:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \approx \arg\max_{\mathbf{e}} \max_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \qquad \textbf{(1)}$$

$\Pr(\mathrm{e}, \mathrm{a} \mid \mathrm{f})$ **is determined through real valued feature functions** $h_i(\mathrm{e}, \mathrm{f}, \mathrm{a}), i = 1 \ldots M$**, and takes the parametric form:**

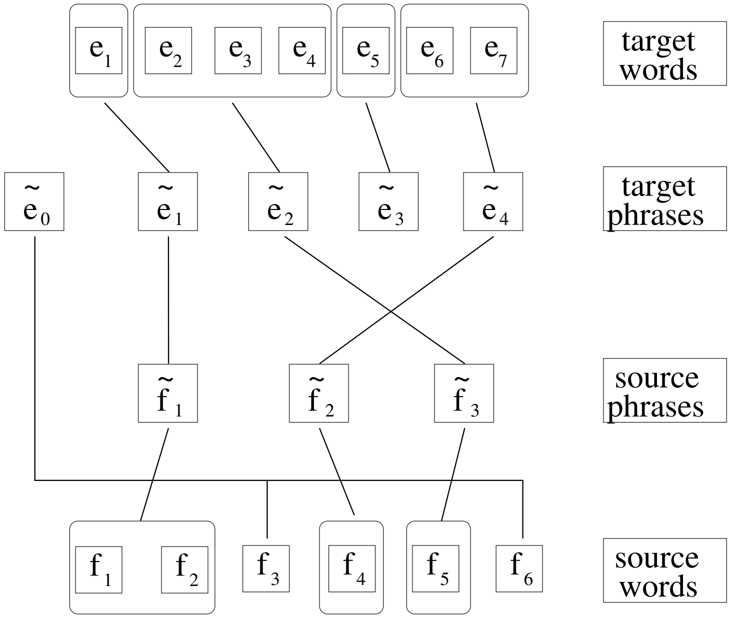
$$p_{\boldsymbol{\lambda}}(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) = \frac{\exp\{\sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}, \mathbf{a})\}}{\sum_{\mathbf{e},a} \exp\{\sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}, \mathbf{a})\}} \qquad \textbf{(2)}$$

Example: feature functions of IBM Model 4:

$$\begin{array}{lcll}
h_1(\mathrm{e}, \mathrm{f}, \mathrm{a}) & = & \log \Pr(\mathrm{e}) & \textbf{(target language model)} \\
h_2(\mathrm{e}, \mathrm{f}, \mathrm{a}) & = & \log \Pr(\boldsymbol{\phi} \mid \mathrm{e}) & \textbf{(fertility model)} \\
h_3(\mathrm{e}, \mathrm{f}, \mathrm{a}) & = & \log \Pr(\boldsymbol{\tau} \mid \mathrm{e}, \boldsymbol{\phi}) & \textbf{(lexicon model)} \\
h_4(\mathrm{e}, \mathrm{f}, \mathrm{a}) & = & \log \Pr(\boldsymbol{\pi} \mid \mathrm{e}, \boldsymbol{\phi}, \boldsymbol{\tau}) & \textbf{(distortion model)}
\end{array}$$

Phrase-based model

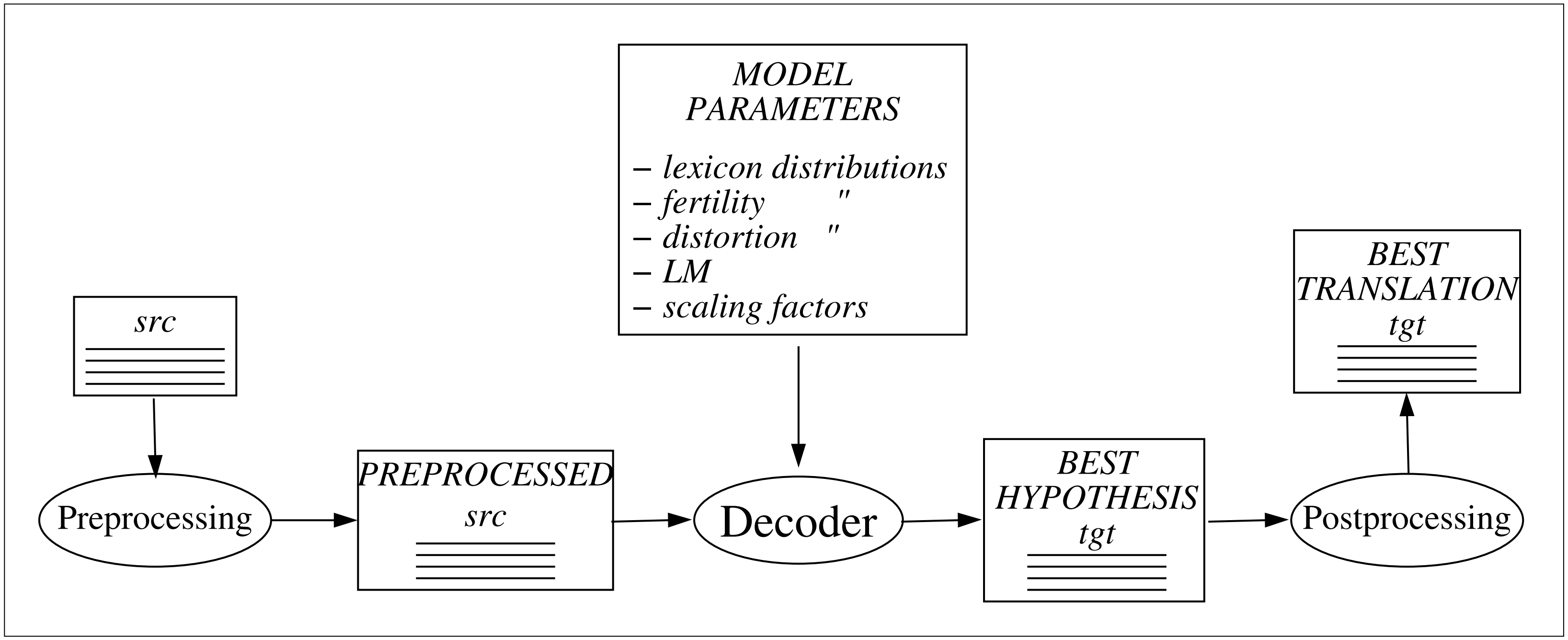
- a *phrase* is a sequence of one or more words (no semantic or syntactic meaning)
- one-to-one correspondence between phrases
- source words may be not translated (into $\tilde{e}_0$)
- insertion of target phrases without translation
- all models at phrase level except language model (at word level)
- frequency-based distributions
- statistics collected from a word alignment (e.g. produced by GIZA++)



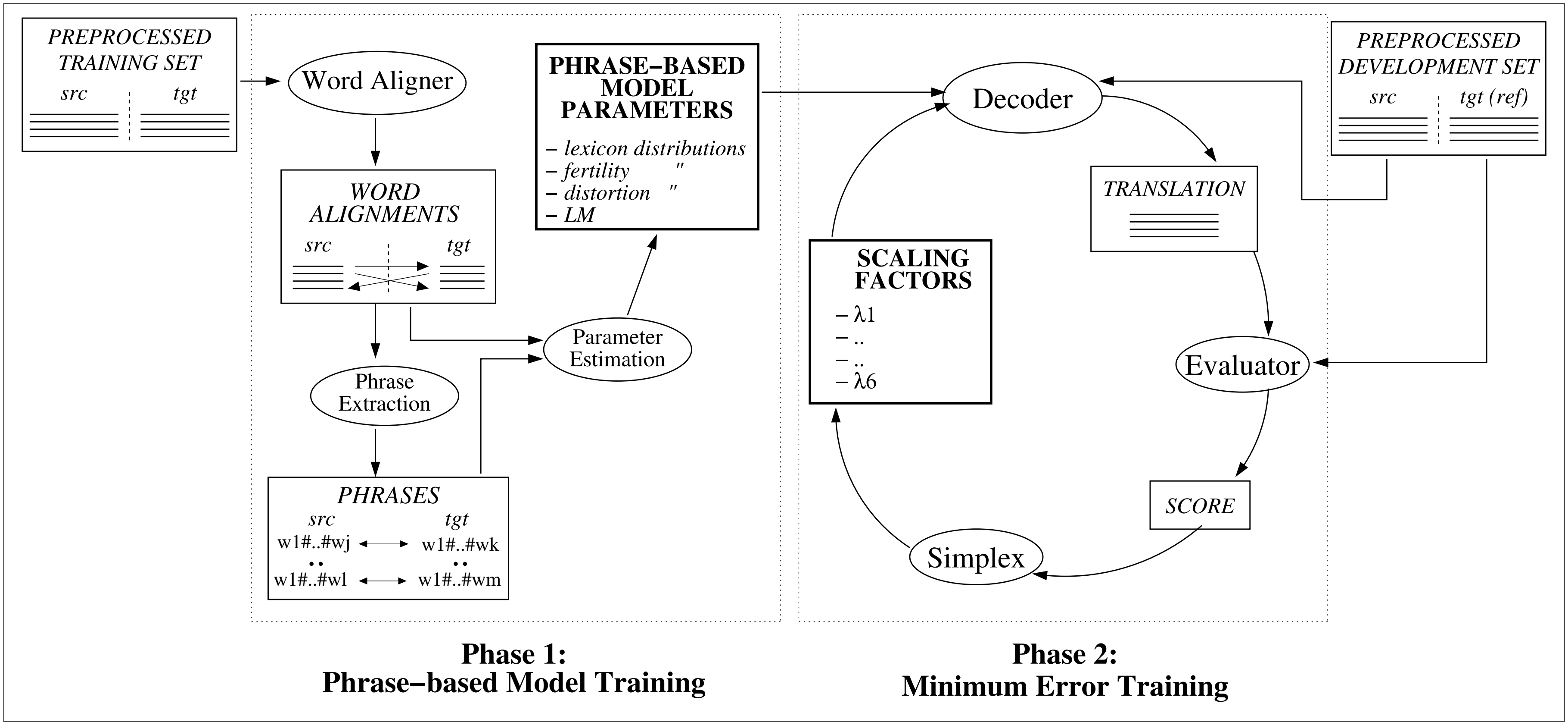
Decoding

- **approximate search criterion:** $\tilde{\mathrm{e}}^* \approx \arg\max_{\tilde{\mathbf{e}}} \ \max_{\mathbf{a}} \sum_i \lambda_i h_i(\tilde{\mathrm{e}}, \mathbf{f}, \mathbf{a})\}$
- **DP-based algorithm**
- **search progresses synchronously along the target string (decisions are taken when generating target phrase)**
- **search ends when all source positions are covered**
- **optimal final theory is chosen among all complete theories**
- **beam search: threshold pruning, histogram pruning**
- **garbaging of theories without extensions**
- **constraints on the length of the source and target phrases**

System Architecture: Run-Time



System Architecture: Training



Phase 1:
Phrase–based Model Training

Phase 2:
Minimum Error Training

Experiments

- Chinese-English track (all the three data conditions)
- no optimization on the post-processing
- BLEU score for data selection and minimum error training

Preprocessing

- tokenization (EN)*
- dp-based Chinese segmentation (CH)*
- rule-based recognition of time and numerical expressions (CH, EN):
  week days, month names, percentages, cardinals, ordinals
- lower case text (EN)
- ignored unknown Chinese words
- split of long sentences (test)

* when needed

Selection of Training Data

System	extra data		BLEU	NIST	MWER	MPER	
name	monolingual	bilingual					
`baseline`			0.3001	7.0157	50.8	41.5	$(\star)$
`lm-btec`	BTEC		0.3509	7.5099	47.2	38.1	$(\star)$
`lm-db1`	BTEC, DB1		0.3466	7.4475	47.6	38.3	
`lm-db2`	BTEC, DB2		0.3460	7.4427	47.1	38.3	
`tm-btec`	BTEC	BTEC	0.4311	8.5336	42.0	33.3	
`tm-db3`	BTEC	BTEC, DB3	0.4574	8.7890	39.7	30.5	$(\star)$

- DB1: news corpora
- DB2: press releases of Hong Kong Special Administrative Region
- DB3: selection of corpora from NIST MT-EVAL 2004 competition (large data condition)

Chinese Segmentation

1. `Supplied`:
   - Chinese segmentation as provided in the supplied training/test corpora
2. `Special`:
   - Chinese segmentation from scratch
   - word-frequency list (7K) extracted from the supplied training corpus
3. `Full`:
   - Chinese segmentation from scratch
   - word-frequency list (44K) provided by LDC

Official Results: Objective Scores

Data Condition	Segmentation		BLEU	NIST	MWER	MPER	
	training	test					
Supplied	`Supplied`	`Supplied`	0.3156	7.1604	53.1	45.3	
	`Special`	`Special`	0.3493	7.0973	50.8	43.0	$(\star)$
Additional	`Supplied`	`Supplied`	0.3499	7.5199	51.0	43.3	
	`Supplied`	`Special`	0.3514	7.3958	49.7	42.0	$(\star)$
	`Supplied`	`Full`	0.3490	6.6185	51.9	44.5	
Unrestricted	`Full`	`Supplied`	0.3774	7.0880	50.0	43.4	
	`Full`	`Special`	0.4118	7.0908	47.7	41.0	
	`Full`	`Full`	0.4409	7.2413	45.7	39.3	$(\star)$

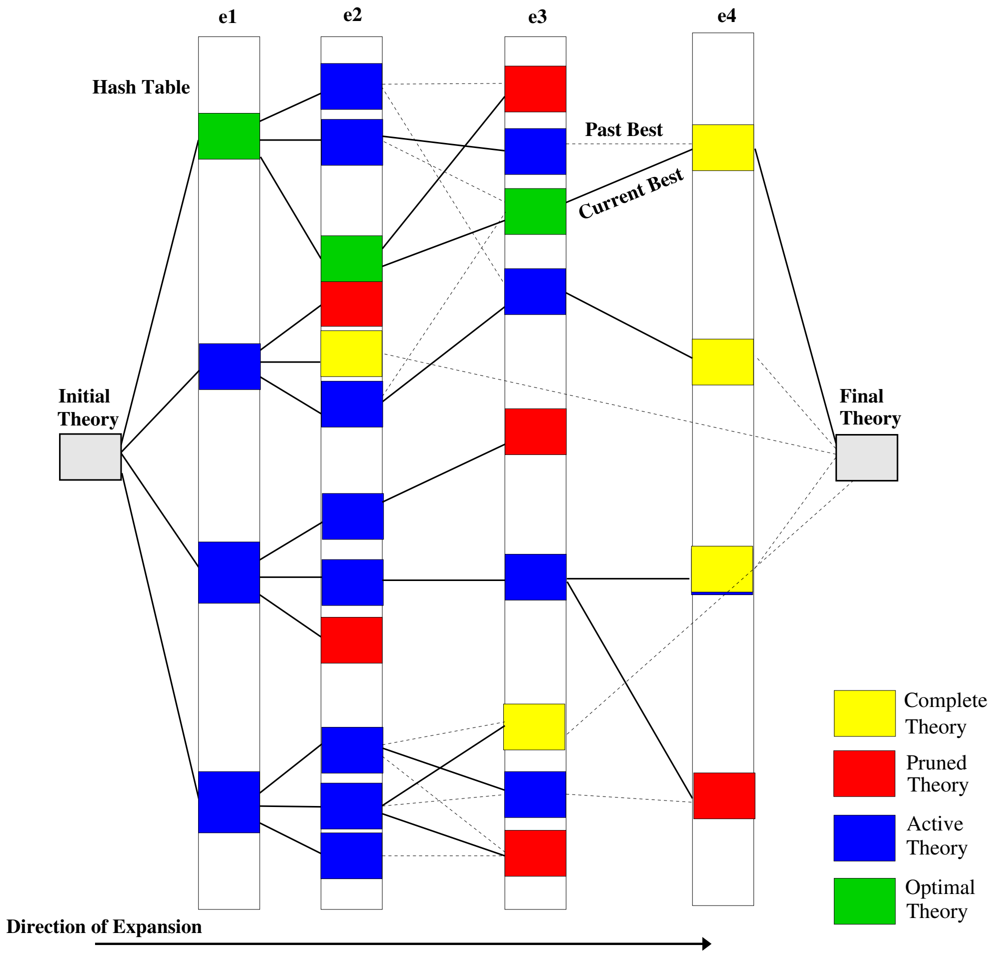
$(\star)$ **marked for subjective evaluation**

Official Results: Subjective Scores

Data Condition	Segmentation		fluency	adequacy
	training	test		
Supplied	`Special`	`Special`	3.120	3.088
Additional	`Supplied`	`Special`	3.256	3.110
Unrestricted	`Full`	`Full`	3.776	3.526

THE END

Decoding: Expansion, Recombination and Pruning


e1
e2
e3
e4
Hash Table
Past Best
Current Best
Initial Theory
Final Theory
Complete Theory
Pruned Theory
Active Theory
Optimal Theory
Direction of Expansion