The ITC-irst SMT System for IWSLT-2005

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

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica

38050 Povo (Trento), Italy

Log-Linear Model Approach to SMT

Maximum Entropy framework for the word-alignment MT approach:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \max_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) = \arg\max_{\mathbf{e}} \max_{\mathbf{a}} \sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}, \mathbf{a})\} \quad \textbf{(1)}$$

where f =source, e =target, a =alignment, and $h_i(\mathrm{e}, \mathbf{f}, \mathrm{a})$ are suitable feature functions.

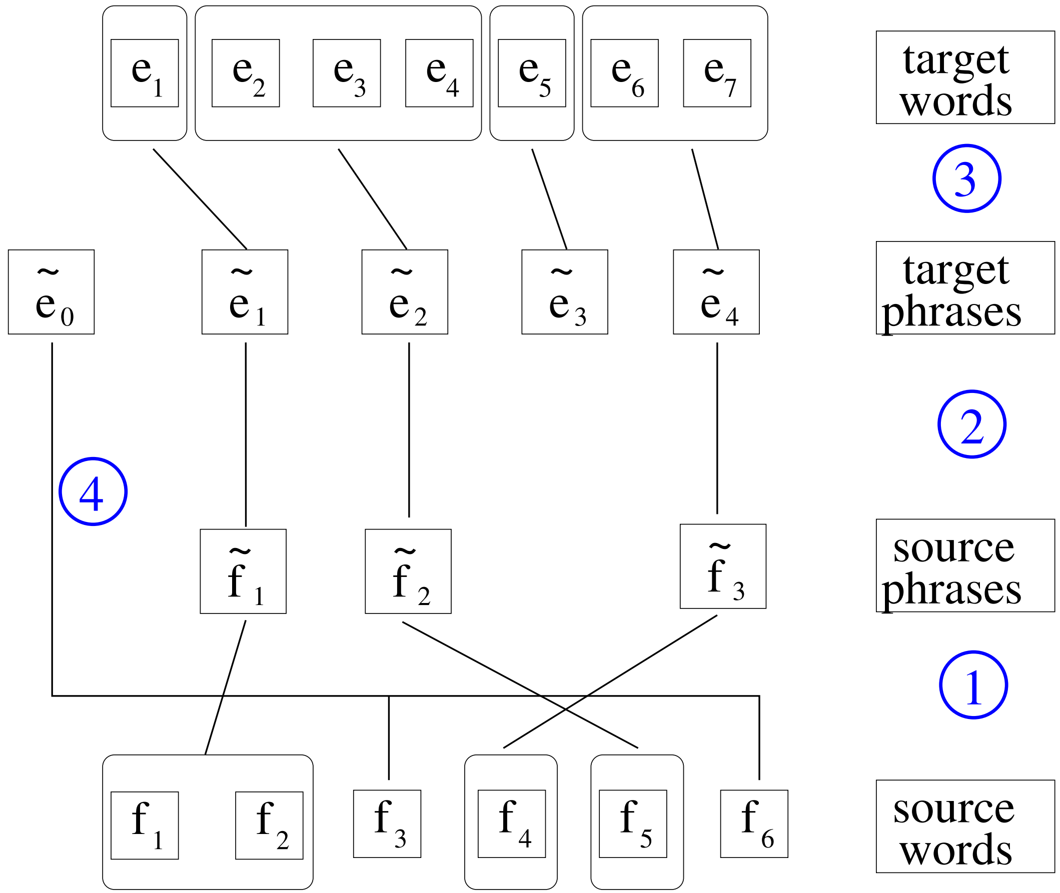
Advantages:

- **directly models the posterior probability (discriminative model)**
- **does not rely on probability factorizations with independence assumptions**
- **is mathematically sound and allows to add any kind of feature function**
- **includes any IBM model as a special case**
- **minimum error training to estimate free parameters (λ_i)**

Phrase-based Model

- A **phrase** is a sequence of one or more words without semantic/syntactic meaning

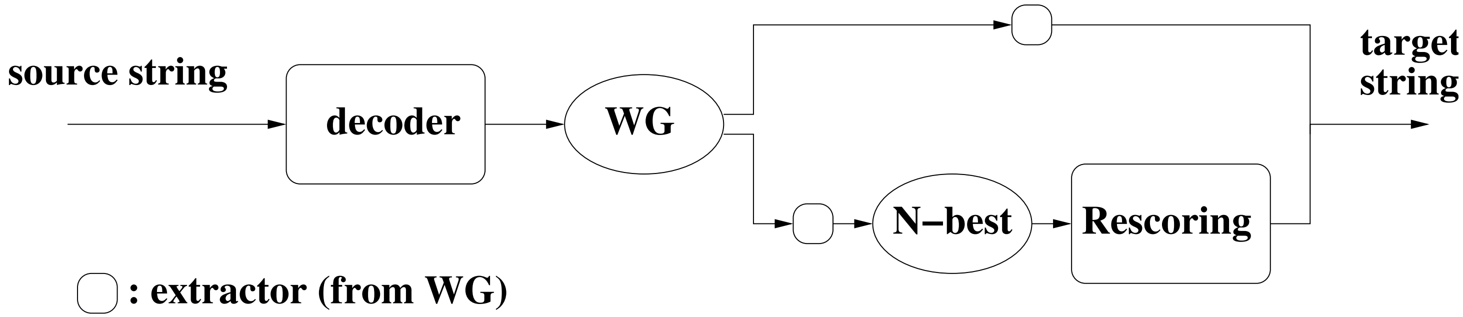
- Generative process:
  1. cover new source positions (distortion)
  2. link to target phrase (fertility,lexicon)
  3. add target phrase (language model)
  4. untranslated words ($\tilde{e}_0$-fertility, lexicon)



Search is over strings of phrases:

$$\tilde{\mathbf{e}}^* = \arg\max_{\tilde{\mathbf{e}}} \max_{\mathbf{a}} \sum_i \lambda_i h_i(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a})\}$$

Two Pass Search Strategy



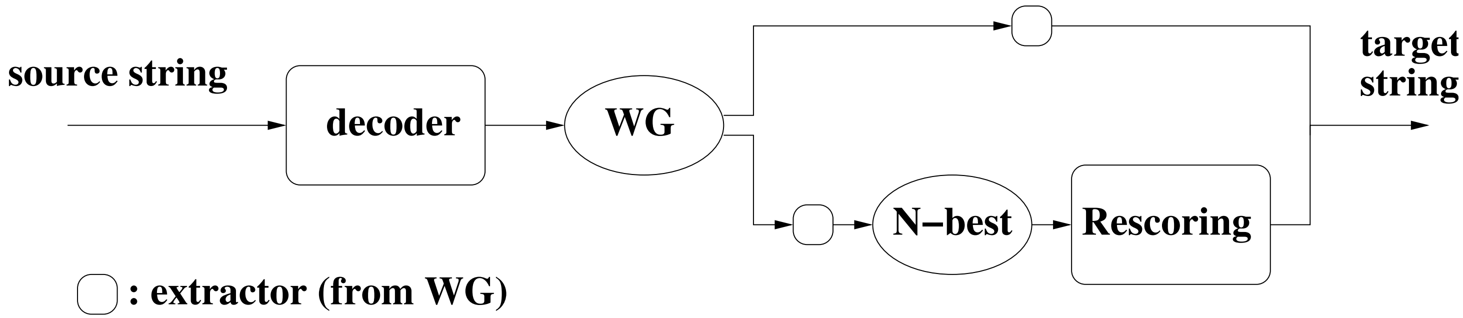
First Pass:

- Log-linear Model
- Dynamic programming algorithm
- Beam search decoder:
  - threshold and histogram pruning
- Non-monotone search constraints
  - max number of vacancies on the left (MVN)
  - max distance from left-most vacancy (MVD)

Second Pass:

- Extraction of 1,000-best
- Log-linear Model
- Re-ranking algorithm

Two Pass Search Strategy



First Pass feature functions:

– Target 3-gram LM

– Fertility model target phrases

– Direct phrase-based lexicon

– Inverse phrase-based lexicon

– Negative distortion

– Positive distortion

– $\tilde{e}_0$ fertility

– $\tilde{e}_0$ permutation

Training of Phrase-based model

Phrase-based model (baseline):

- **Word-alignment: union of direct and inverse IBM alignments (GIZA++, $1^5H^53^44^45^4$)**
- **Phrase-extraction: max length 8, filtering (length or punctuation mismatches)**
- **Feature estimation: lexicon, fertility models (... by freq smoothing ...)**
- **Monotone search: MVD=0**

Improvements by exploiting Competitive Linking Algorithm (Melamed, 2000):

- **CLA translation lexicon added to data before word-alignment**
- **CLA word-alignments added to IBM word alignments before phrase-extraction**
- **Re-segmented Chi/Jap data added to training data before word-alignment (in-house tool)**

Experimental Results: First Pass

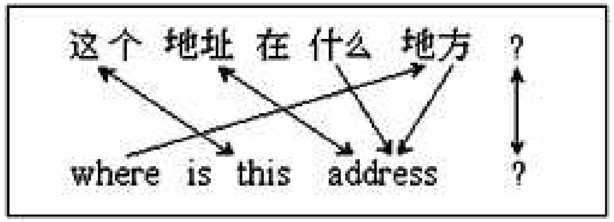
- Task: Supplied Data Condition
- Lang: Chinese,Japanese, Arabic
- Test set: IWSLT 2004
- Dev set: CSTAR 2003
- BLEU%:no-case with punctuation
- No weight optimization
- Non-monotone search:
  - MVD=4 MVN=3 Arabic
  - MVD=6 MVN=5 Chinese
  - MVD=7 MVN=6 Japanese

System	Chi2Eng	Jap2Eng	Ara2Eng
baseline	35.82	33.82	51.01
+CLA translation lexicon	36.28	35.78	52.84
+CLA alignments	37.59	38.77	54.14
+re-segmented data	38.29	38.97	–
+chunked data	–	39.59	–
+non-monotone search	42.51	44.66	56.40

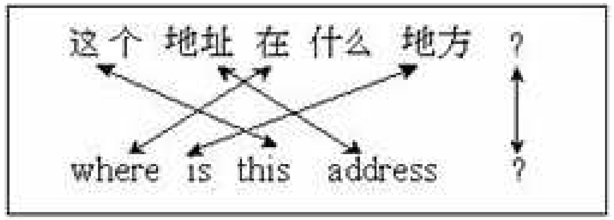
CLA alignments vs. IBM Alignments

- IBM alignments are many-to-one
- CLA alignments are one-to-one
- CLA alignments have higher precision
- CLA alignments allow for more phrase-pairs

IBM alignments (direct and inverse):

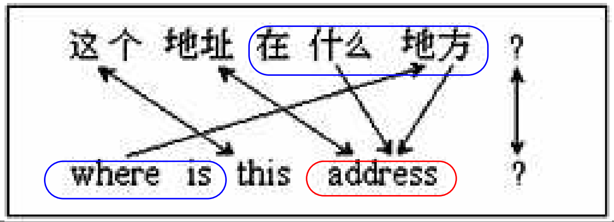


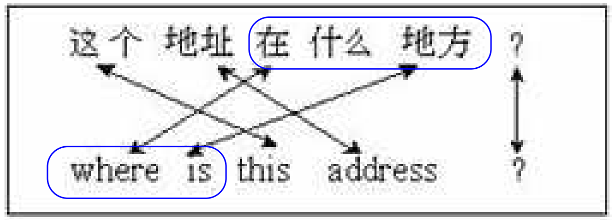
CLA alignment:



Despite past work (Och & Ney, 2003) showed that quality of CLA alignments is poorer than for IBM Model 1, we found that such alignments work indeed well for phrase-based SMT.

Phrase extraction from IBM and CLA alignments

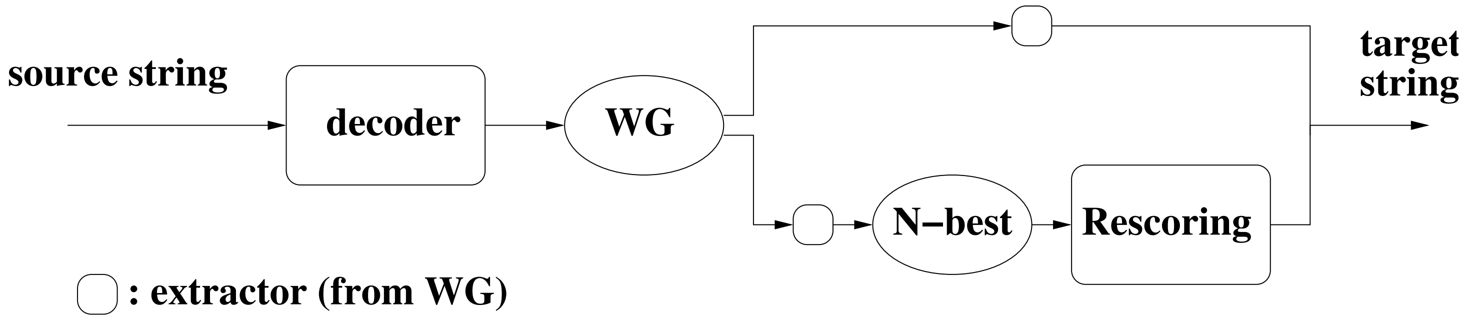




IBM Alignment		CLA Alignment	
NULL_	is	这个	this
这个	this	地址	address
这个	is this	在	where
地址	address	什么	NULL_
在	NULL_	地方	is
什么	address	?	?
地方	address	这个 地址	this address
?	?	在 什么	where
这个 地址 在 什么 地方	where is this address	什么 地方	is
这个 地址 在 什么 地方 ?	where is this address ?	在 什么 地方	where is
		这个 地址 在 什么 地方	where is this address
		这个 地址 在 什么 地方 ?	where is this address ?

In this real example, the CLA alignment allows to extract the useful phrase "where is".

Two Pass Search Strategy



Second Pass feature functions:

– IBM model 1 lexicon score

– IBM model 3 lexicon score

– CLA lexicon score

– Question feature

– Frequency of n-grams within n-best

– ratio of target source lengths

– 2-gram target LM

– 4-gram target LM

– 5-gram target LM

New Feature Functions in Re-scoring

The following statistics are computed on each entry of the 1000-best list:

- **CLA alignment score**

  Integrates the CLA associative score over all possible word alignments between source and target, similarly to how is done for IBM Model 1 re-scoring

- **Question tag**

  Triggers a binary feature when the string ends with a question mark and starts with one of the following words: what, which, who, when, how, do, did, ...

- **N-gram frequency**

  Counts the frequencies of its n-grams (n=1,2,3,4) within the full n-best list and sums them up according to a linear combination.

Experimental Results: Re-scoring Stage

- Task: Supplied Data Condition
- Lang: Chinese,Japanese, Arabic
- BLEU%:no-case with punctuation
- Test set: IWSLT 2004
- Dev set: CSTAR 2003
- Optimization: BLEU% + 4 * NIST
- N-best 1000

System	Chi2Eng	Jap2Eng	Ara2Eng
Decoder	42.51	44.66	56.40
IBM model-1	42.31	44.48	56.00
IBM model-3	41.53	44.97	56.16
CLA score	42.42	45.20	56.31
question tag	42.81	45.83	56.66
n-grams	43.71	46.19	56.89
target length	41.11	41.00	50.87
2-grams LM	44.06	45.34	56.07
4-grams LM	45.88	45.51	56.72
5-grams LM	45.72	45.81	56.61
+all features	47.99	51.01	57.94

Conclusions

Main performance improvements came from:

- **Integration of IBM and CLA word-alignments at different levels:**
  - **Translation lexicon used to constrain IBM alignments**
  - **Phrase-extraction performed on both CLA and IBM word-alignments**
- **Use of multiple word segmentations for Chinese,Japanese**
- **New feature functions used for n-best re-scoring:**
  - **Associative score from CLA**
  - **Frequency of n-grams in n-best list**
  - **High order language models (4-gram 5-gram)**
- **Optimization of non-monotone search constraints**

The End ... Thank You!