Improving Phrase-Based Statistical Translation through Combination of Word Alignments

B. Chen, 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:

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

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

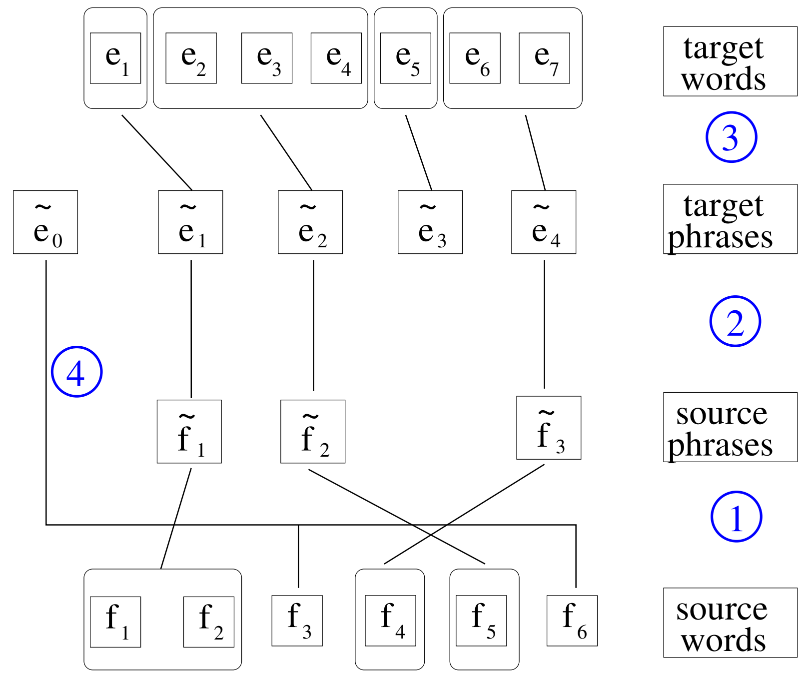
Advantages:

- **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

- Translation 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:

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

Search Strategy

- 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)

- Feature Functions:
  - Target 4-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

Competitive Linking Algorithm (Melamed, 2000)

- **Under the one-to-one assumption**
- **An association score is computed for every possible word pair – a log-linear combination of two probabilities (Kraif & Chen, 2004):**
  - **1) word pairs co-occurrence**
  - **2) word pairs position difference**

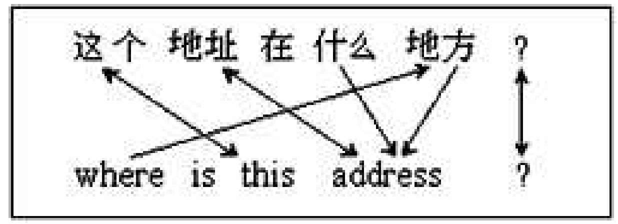
$$S(f_j, e_i) \;=\; -\log P_{cooc}(f_j, e_i) + k \log P_{pos}(dist(j, i))$$

- **Apply a greedy algorithm to select the best word-alignments**
  - **1) Sort all the word pairs $Cand$ in descending order of the association score;**
  - **2) The best scoring pair is restored in $link$ and removed from $Cand$**
  - **3) All the competing word pairs are removed from $Cand$;**
  - **4) Return to 1), until $Cand$ is empty.**

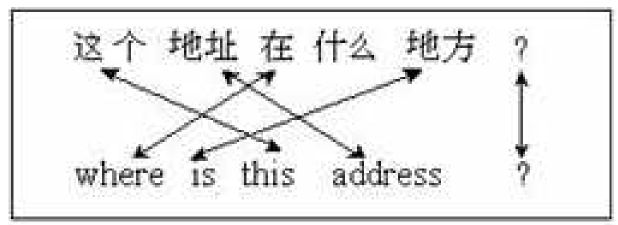
CLA alignments vs. IBM Alignments

- IBM alignments are many-to-one
- CLA alignments are one-to-one
- CLA alignments have higher precision

IBM alignments (a & b):



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.

Training of Phrase-based model

Phrase-based model:

- Word-alignment:
  - 1) IBM Union: $a \cup b$
  - 2) CLA: Competitive Linking Algorithm
  - 3) IBM Inter.: $a \cap b$ with expansion (Och etc., 1999)
- Phrase max length: 8
- Feature estimation: lexicon, fertility models (... by freq smoothing ...)
- Non-monotone search

Improved by CLA:

– 4) Inter+CLA: phrase-pairs obtained from 2) and 3) are joined.

Experiments

- Task:
  - Chinese, Japanese, Arabic: IWSLT 2005 Supplied Data Condition (20K sentence-pairs)
  - Italian: equivalent test-suite from C-STAR Consortium
  - Chinese and Italian: extended It-En and Ch-En up to 60K and 160K sentence-pairs
- Dev set: IWSLT 2004
- Test set: CSTAR 2003
- BLEU% & NIST: no-case, with punctuation
- Weight optimization
- Non-monotone search:
  - MVD=4 Arabic, Chinese, Japanese
  - MVD=2 Italian

Statisticals of Experiments Data (1)

Statistics of training, development and testing data used for the IWSLT 2005 supplied data condition. For Italian-English a comparable set was collected.

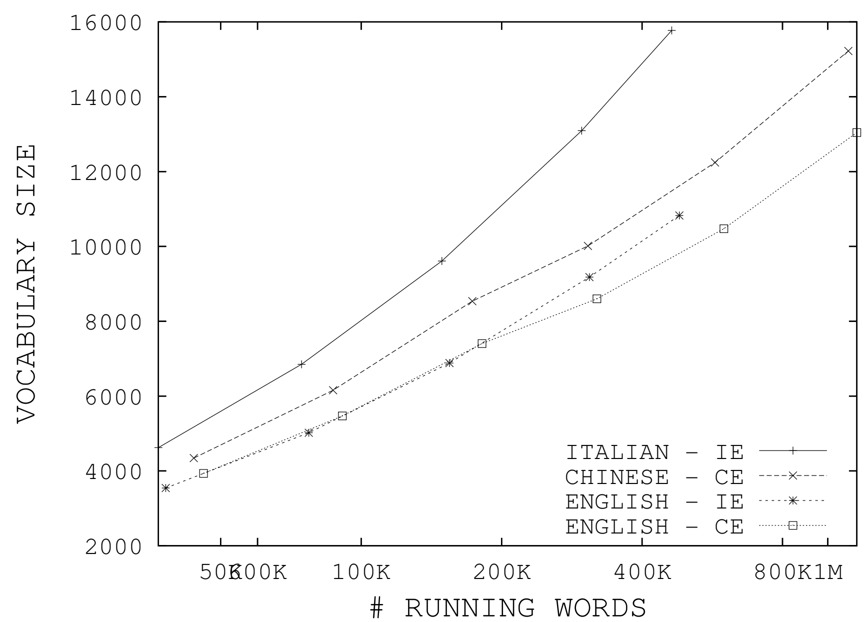
		IWSLT 2005				Italian-English	
		Chinese	Arabic	Japanese	English	Italian	English
Train Data	Sentences	20,000				20,000	
	Running words	173K	171K	159K	181K	149K	155K
	Vocabulary	8,536	9,251	18,150	7,348	9,611	6,885
Dev. Data	Sentences	500			500 $\times$ 16	100	100 $\times$ 16
	Running words	3,860	3,538	3,359	64,884	788	14,001
Test Data	Sentences	506			506 $\times$ 16	506	506 $\times$ 16
	Running words	3,514	3,531	3,259	65,616	3,574	65,615

Statisticals of Experiments Data (2)

Statistics of extended BTEC data

Training Data	Chinese	English
Sentences	160,000	
Running words	1,106K	1,154K
Vocabulary	15,222	13,043
Training Data	**Italian**	**English**
Sentences	60,000	
Running words	463K	480K
Vocabulary	15,775	10,828

Vocabulary growth in the extended BTEC data



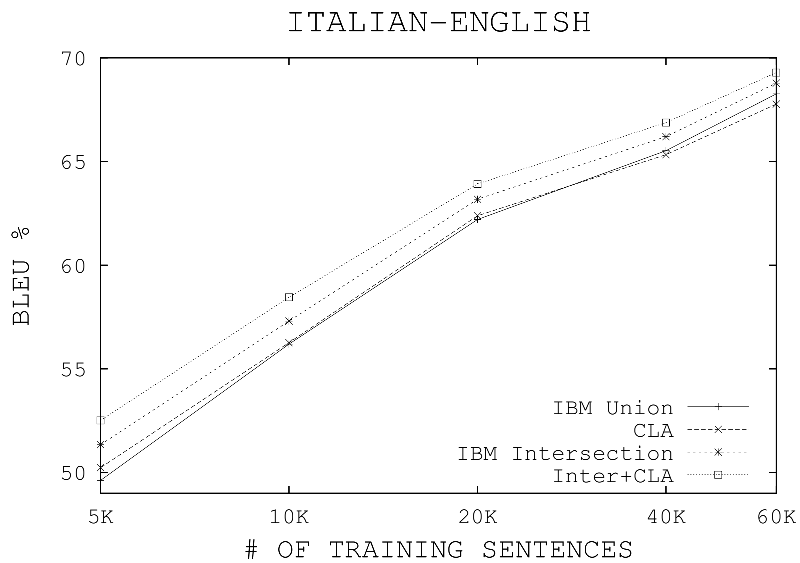
Experiment Results (1)

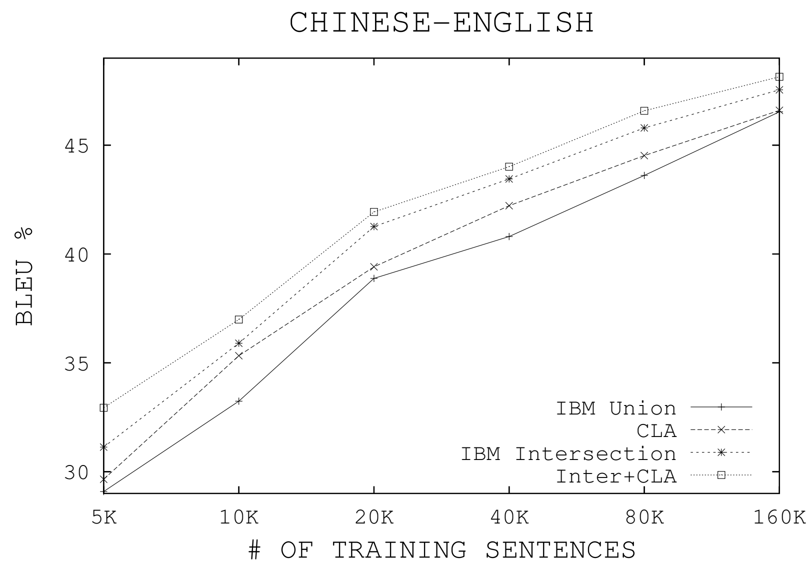
BLEU% scores and NIST scores for different Language pairs in the IWSLT 2005 supplied data condition

Language	Chinese		Janpanese		Arabic		Italian	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
IBM Union	38.88	7.411	42.52	7.731	58.23	8.880	62.20	9.846
CLA	39.41	7.457	45.96	7.770	57.26	8.977	62.38	9.822
IBM Inter.	41.26	7.387	46.59	7.778	59.05	8.925	63.18	9.842
Inter+CLA	41.93	7.492	47.76	7.858	59.79	9.191	63.92	9.853

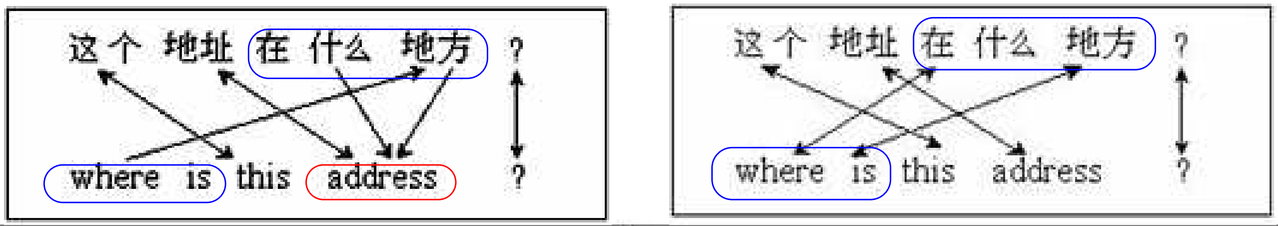
Experiment Results (2)

Performance of training modalities against increasing amounts of training data





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".

Conclusions

- Comparison of SMT performance on three word-alignments:
  - IBM Union word-alignments
  - CLA word-alignments
  - IBM Intersection word-alignments
- Integration of IBM and CLA word-alignments gives consistent improvements

The End ... Thank You!