

# Discriminative Alignment Training without Annotated Data for Machine Translation

*Patrik Lambert, Rafael E. Banchs and Josep M. Crego*

TALP Research Center  
Jordi Girona Salgado, 1-3  
08034 Barcelona, Spain

NAACL 2007, Rochester, NY

- 1 Introduction
- 2 Bilingual Word Aligner
- 3 Experimental Work
- 4 Conclusions

# Introduction

- In IBM models: word alignment hidden variable of translation model
- In most current SMT systems, word alignment and translation model training **separated** in two steps.
- Difficulties of tuning (word) alignment for Machine Translation (MT):

practical difficulties depending on alignment system:

- generative models: not easily adaptable, large computational resources
- discriminative approach: need of human reference

main difficulty: absence of adequate metric

- bad correlation between alignment quality (AER) and translation quality
- new metrics proposed [Fraser and Marcu 2006; Ayan and Dorr 2006]: more informative for translation units extraction, but **indirect** method

## Our Method

Tuning alignment parameters **directly** in a Minimum translation Error Training scheme: use automated translation metrics as minimization criterion.

Discriminative framework for word alignment with log-linear models [Moore 2005]: seems adequate

- easy to incorporate new features
- no observation sequence (*i.e.* manually annotated data) needed, just a score (*e.g.* MT metric)
- not globally optimal solution to the alignment problem (beam search).  
however: better alignments do not imply better translations

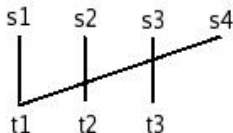
# BIA: Bilingual word Aligner

- Beam-search decoder minimizing cost of linear combination of models [Moore 2005]
- Some features designed to suit N-gram based SMT system
- Particularity of N-gram SMT: translation model, based on N-gram language model of bilingual units (*tuples*).
- Two issues can be dealt with at the alignment stage:
  - only one monotonic segmentation of each sentence pair is performed  
⇒ long reorderings produce long and sparse tuples
  - tuples with NULL source sides cannot be allowed ⇒ process alignments to attach any unlinked target word to its precedent or its following word

## Aligner Features

We took these two issues into account to implement the following features:

- distinct source and target unlinked word penalties
- embedded word position penalty: penalizes situations like this one:



- link bonus: because N-gram model prefers higher recall alignments

Other features:

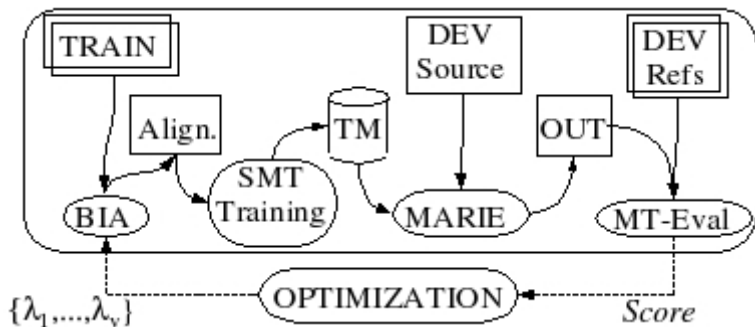
- two distortion features: one sums the number of crossing links, and the other one sums the amplitude of crossing links
- $\phi^2$  score [Gale and Church 1991] as word association model, and as POS-tag association model.

# Data set

- IWSLT06 Mandarin→English.
- Training: 46000 sentences of 7 words in average
- Development: 489 sentences of 11.2 words in average, with 7 references
- Test: 500 sentences of 6 words in average, with 16 references

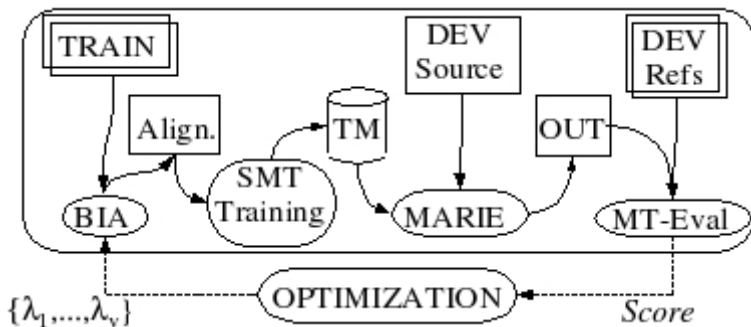
## Procedure

- Optimal coefficients were estimated with the following procedure:



## Procedure

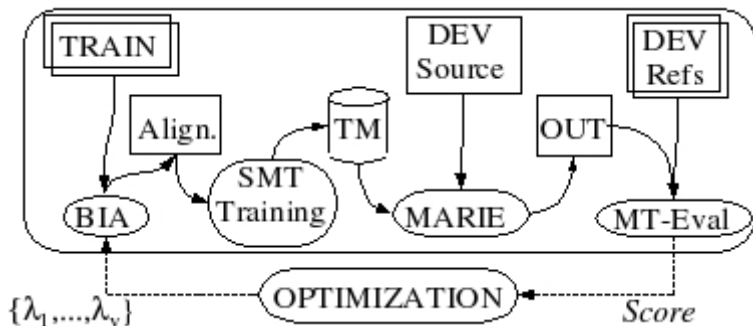
- Optimal coefficients were estimated with the following procedure:



- corpus was aligned with optimal coefficients
  - translation units
  - full SMT system (with TM + target language model, word bonus model and two lexical models).

## Procedure

- Optimal coefficients were estimated with the following procedure:



- corpus was aligned with optimal coefficients
  - translation units
  - full SMT system (with TM + target language model, word bonus model and two lexical models).
- GIZA++ (50 cl, 1-4 H-5 4-4) → translation units → full SMT system

## Results

- in full SMT system, model weights were optimized by minimum error training (MET), and test corpus was translated
- MET performed 3 times: average (standard deviation) are shown

System	BLEU	NIST	PER	WER
bia (BLEU)	44.8 (0.4)	<b>9.00</b> (0.04)	35.7 (0.07)	43.8 (0.09)
bia (BLEU+4NIST)	<b>47.0</b> (1.5)	8.83 (0.4)	<b>32.9</b> (0.8)	<b>40.9</b> (0.5)
bia (NIST)	44.8 (0.1)	8.55 (0.14)	<b>33.0</b> (0.2)	41.4 (0.5)

- large variability implied by different sets of BIA aligner coefficients

## Results

- in full SMT system, model weights were optimized by minimum error training (MET), and test corpus was translated
- MET performed 3 times: average (standard deviation) are shown

System	BLEU	NIST	PER	WER
giza++ union	42.7 (1.1)	8.82 (0.07)	34.7 (0.2)	43.7 (0.4)
bia (BLEU)	44.8 (0.4)	<b>9.00</b> (0.04)	35.7 (0.07)	43.8 (0.09)
bia (BLEU+4NIST)	<b>47.0</b> (1.5)	8.83 (0.4)	<b>32.9</b> (0.8)	<b>40.9</b> (0.5)
bia (NIST)	44.8 (0.1)	8.55 (0.14)	<b>33.0</b> (0.2)	41.4 (0.5)

- large variability implied by different sets of BIA aligner coefficients
- better scores in general for BIA

## Results

- in full SMT system, model weights were optimized by minimum error training (MET), and test corpus was translated
- MET performed 3 times: average (standard deviation) are shown

System	BLEU	NIST	PER	WER
giza++ union	42.7 (1.1)	8.82 (0.07)	34.7 (0.2)	43.7 (0.4)
bia (BLEU)	44.8 (0.4)	<b>9.00</b> (0.04)	35.7 (0.07)	43.8 (0.09)
<b>bia (BLEU+4NIST)</b>	<b>47.0</b> (1.5)	<b>8.83</b> (0.4)	<b>32.9</b> (0.8)	<b>40.9</b> (0.5)
bia (NIST)	44.8 (0.1)	8.55 (0.14)	<b>33.0</b> (0.2)	41.4 (0.5)

- large variability implied by different sets of BIA aligner coefficients
- better scores in general for BIA
- one configuration gives really better results

## Results

- in full SMT system, model weights were optimized by minimum error training (MET), and test corpus was translated
- MET performed 3 times: average (standard deviation) are shown

System	BLEU	NIST	PER	WER
giza++ union	42.7 (1.1)	8.82 (0.07)	34.7 (0.2)	43.7 (0.4)
bia (BLEU)	44.8 (0.4)	<b>9.00</b> (0.04)	35.7 (0.07)	43.8 (0.09)
bia (BLEU+4NIST)	<b>47.0</b> (1.5)	8.83 (0.4)	<b>32.9</b> (0.8)	<b>40.9</b> (0.5)
bia (NIST)	44.8 (0.1)	8.55 (0.14)	<b>33.0</b> (0.2)	41.4 (0.5)

- large variability implied by different sets of BIA aligner coefficients
- better scores in general for BIA
- one configuration gives really better results
- the more we use NIST in optimization, the worse is NIST score in test. Doesn't occur in development results. Due to size of sentences in development and test (nearly a factor 2 difference) ?

# Conclusions

- novel framework for discriminative training of alignment models with automated translation metrics as optimization criterion.
- SMT systems trained with this framework's alignments outperformed the ones trained with Giza++ alignment combinations.
- Further work:
  - improve this basic first version of aligner, especially the association score model (e.g. add discount factors or add more association score types, or dictionaries). Add a fertility model.
  - during alignment coefficient optimization, only the translation model is used  $\Rightarrow$  consider using various SMT features (as would be required for a phrase-based SMT system).
  - approach difficult to apply to a large corpus (whole corpus must be aligned at each iteration)  $\Rightarrow$  determine whether alignment parameters trained on a part of the corpus are valid for the whole corpus.
  - tune GIZA++ parameters in the same way as for BIA parameters.