The TALP Ngram-based SMT System for IWSLT 2007

Patrik Lambert, **Marta R. Costa-jussà**, *Josep M. Crego*,
Maxim Khalilov, *José B. Mariño*,
Rafael Banchs, *José A.R. Fonollosa and Holger Schwenk*[1]

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

[1] LIMSI-CNRS, BP 133
91403 Orsay Cedex
schwenk@limsi.fr

IWSLT 2007, Trento

Translation Model

The best translation hypothesis $\mathbf{T}$, for a given source sentence $\mathbf{S}$, is that which maximizes a log-linear combination of feature functions:

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}} \sum_m \lambda_m h_m(\mathbf{T}, \mathbf{S})$$

- Translation Model:
  N-gram language model of bilingual units (tuples)

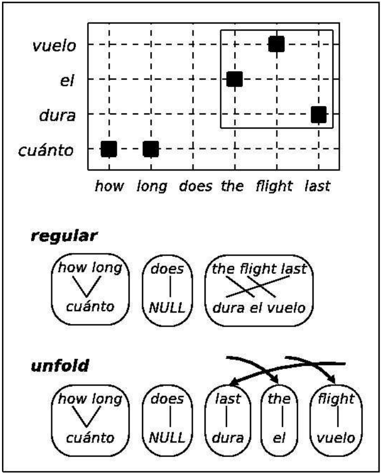
$$p(\mathbf{T}, \mathbf{S}) \approx \prod_n p((t,s)_n | (t,s)_{n-N+1}, \ldots, (t,s)_{n-1})$$

Tuple extraction

Tuples are extracted from word alignment

- A unique and monotonic segmentation of each sentence is produced.
- No word in a tuple is aligned to words outside of it
- No smaller tuples can be extracted without violating the previous constraints

Tuple extraction example



Unfolding produces a different bilingual n-gram model with reordered source words.

Additional feature functions

Additional feature functions:

- Target language model
- POS target language model
- Word bonus model, giving a bonus proportional to the number of target words.
- Source-to-target and target-to-source lexicon models, which compute a lexical weight for each tuple, using IBM model 1 translation probabilities

Alignment Minimum Translation-Error Training

Our Method

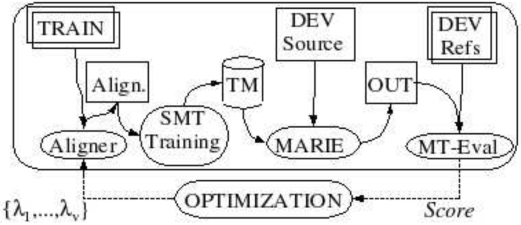
Tuning alignment parameters directly in a Minimum translation Error Training scheme: use automated translation metrics as minimization criterion.

Alignment optimization parameters chosen for GIZA++:

- Smoothing factors for models HMM, IBM3 and IBM4
- The probability for the empty word
- Deficient distorsion for the empty word

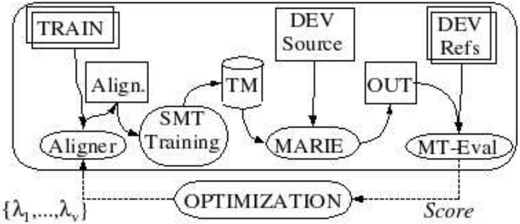
Procedure

- Optimal coefficients were estimated with the following procedure:



Procedure

- Optimal coefficients were estimated with the following procedure:



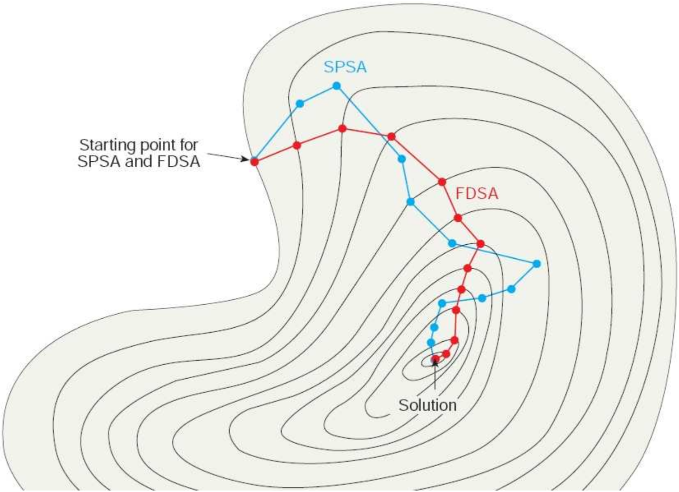
- SMT system with TM model (bilingual language model)

Simultaneous Perturbation Stochastic Approximation

- The SPSA method [J. Spall, 1992] is based on a gradient approximation which requires only two evaluations of the objective function, regardless of the dimension of the optimisation problem.
- SPSA procedure is in the general recursive stochastic approximation form:

$$\hat{\lambda}_{k+1} = \hat{\lambda}_k - \mathbf{a}_k \hat{\mathbf{g}}_k(\hat{\lambda}_k)$$

$\hat{\mathbf{g}}_k(\hat{\lambda}_k)$: estimate of the gradient $\mathbf{g}(\lambda) \equiv \partial E / \partial \lambda$ at iterate k



The simultaneous approximation causes deviations of the search path.
These deviations are averaged out in reaching a solution.

Optimization schemes

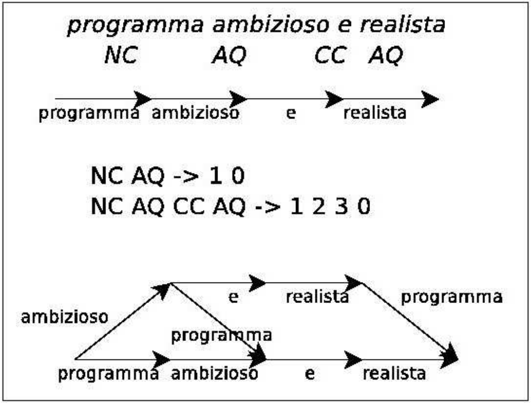
Concept	Procedure	Optimized parameters
Dev translated in each iteration	*Alignment Minimum Translation-Error Training*	*Align. smoothing factors*

Optimization schemes

Concept	Procedure	Optimized parameters
Dev translated in each iteration	*Alignment Minimum Translation-Error Training*	*Align. smoothing factors*
Nbest-list produced by the decoder	*(Double-loop) Minimum Error Training*	*Translation feature functions*

Reordering patterns

Use a set of rewrite rules for Part-Of-Speech sequences to extend the monotonic search graph with reordering hypotheses



Neural Network Language Model

The basic idea of the neural network LM is to project the word indexes onto a continuous space and to use a probability estimator operating on this space.

- The resulting probability functions are smooth functions of the word representation $\longrightarrow$ better generalization to unknown n-grams can be expected.
- A neural network $\longrightarrow$ simultaneously learns the projection of the words onto the continuous space and estimates the n-gram probabilities.

Neural Network Language Model

The basic idea of the neural network LM is to project the word indexes onto a continuous space and to use a probability estimator operating on this space.

- The resulting probability functions are smooth functions of the word representation → better generalization to unknown *n*-grams can be expected.
- A neural network → simultaneously learns the projection of the words onto the continuous space and estimates the *n*-gram probabilities.

The LM posterior probabilities are "interpolated" for any possible context of length *n*-1 instead of backing-off to shorter contexts.

Data Preprocessing

- Training sentences were split by using final dots on the bilingual text

Data Preprocessing

- Training sentences were split by using final dots on the bilingual text
- **Arabic**
  → MADA+TOKAN system for disambiguation and tokenization.
  → This tool produces POS tags on all taggable tokens.

Data Preprocessing

- Training sentences were split by using final dots on the bilingual text
- **Arabic**
  → MADA+TOKAN system for disambiguation and tokenization.
  → This tool produces POS tags on all taggable tokens.
- **Chinese**
  → Resegmentation using ICTCLAS
  → POS tagging using the freely available Stanford Parser

Data Preprocessing

- Training sentences were split by using final dots on the bilingual text
- **Arabic**
  $\rightarrow$ MADA+TOKAN system for disambiguation and tokenization.
  $\rightarrow$ This tool produces POS tags on all taggable tokens.
- **Chinese**
  $\rightarrow$ Resegmentation using ICTCLAS
  $\rightarrow$ POS tagging using the freely available Stanford Parser
- **English**
  $\rightarrow$ Part-Of-Speech tagging *TnT* tagger.
  $\rightarrow$ For alignment purpose only (of the ZhEn system), the Snowball stemmer.

Experimental Settings

- Alignment parameters
  - running 5, 5, 3 and 3 iterations of models 1, HMM, 3 and 4,
  - using English stems and 50 classes,
  - taking the union of source-target and target-source alignments.
- Decoding parameters
  - the beam search was set to 50,
  - no reordering limit in search (all paths present in the input reordering graph are considered).
- Rescoring
  - incorporation of the NNLM into the SMT system was done using 1000-best lists.

Internal Experiments Summary

	dev (dev4)	test (dev5)		
	$\frac{1}{2}$(BLEU+METEOR)	BLEU	NIST	METEOR
Chinese→English				
baseline	0.340	0.186	5.84	0.487
giza++ MET	0.349	0.190	5.97	0.490
giza++ MET+NNLM	0.350	0.205	6.06	0.496

Table: *Internal translation results for IWSLT 2007 Chinese-English task. MET refers to alignment tuning with Minimum (translation) Error Training. NNLM refers to rescoring a translation N-best list with a continuous space target language model.*

Participation in the IWSLT 2007 Evaluation

	UPC	Best	Rank
AE ASR Primary	0.4445	0.4445	1/11
AE Clean Primary	0.4804	0.4923	3/11
CE Clean Primary	0.2991	0.4077	11/15
CE Clean Primary + NNLM	0.2920	0.4077	-

Table: *Official translation results (BLEU scores) for IWSLT 2007 Chinese-English and Arabic-English tasks. Next to our system's score, we indicated the Best system's score. For the primary runs, we also indicated the rank of our system among all primary runs.*

1. TALP Ngram-based Translation System
2. Alignment Minimum Translation-Error Training
3. Simultaneous Perturbation Stochastic Approximation method
4. Word ordering strategies
5. Neural Network Language Model
6. Experiments
7. **Conclusions and Further Work**

Conclusions and further work

1. The optimization of alignment parameters allows to improve translation when using the Alignment Minimum Translation-Error Training.

Conclusions and further work

1. The optimization of alignment parameters allows to improve translation when using the Alignment Minimum Translation-Error Training.
2. The NNLM obtained an improvement of 1.5 Bleu in the internal set.

Conclusions and further work

1. The optimization of alignment parameters allows to improve translation when using the Alignment Minimum Translation-Error Training.
2. The NNLM obtained an improvement of 1.5 Bleu in the internal set.
3. Our system was ranked 1st in the Arabic-English task. It was not very competitive in the Chinese-English task.

Thanks

Grazie a tutti

{lambert, mruiz, jmcrego, khalilov, canton, rbanchs, adrian}@gps.tsc.upc.edu
schwenk@limsi.fr