The TALP Ngram-based SMT System for IWSLT 2006

Josep M. Crego, Adrià de Gispert, **Patrik Lambert**, *Maxim Khalilov, Marta R. Costa-jussà, José B. Mariño, Rafael Banchs and José A.R. Fonollosa*

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

IWSLT 2006, Kyoto

1. TALP Ngram-based Translation System
2. Tuple segmentation strategies
3. Word ordering strategies
4. Experiments
5. Conclusions and Further Work

Participation in the IWSLT 2006 Evaluation

- Tasks
  - Arabic to English
  - Chinese to English
  - Italian to English
  - Japanese to English
- System
  - TALP-tuples (TALP Ngram-based SMT system)

Translation Model

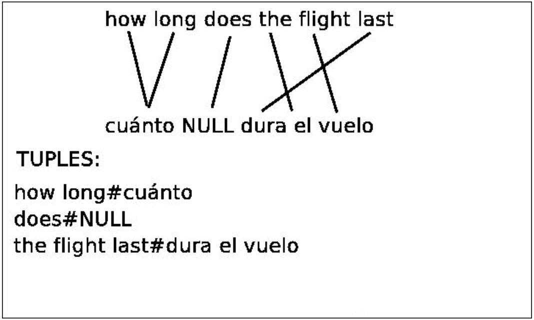
The best translation hypothesis **T**, for a given source sentence **S**, is that which maximises a log-linear combination of 5 models:

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

- 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, monotonous segmentation of each sentence pair 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

Additional feature functions

Additional feature functions:

- 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

Decoding

Decoding:

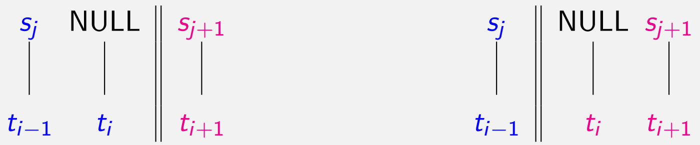
- freely available MARIE decoder [Crego *et al.*, 2005]
  (beam search with hypothesis recombination, threshold and histogram pruning)
- no rescoring module (1-best output used)
- monotone and reordered search

Feature function weights optimization: Downhill Simplex Method

NULL-source tuples

Tuple extraction algorithm defines a unique set of tuples except whenever the resulting tuple contains no source word (NULL-source tuple).

These units cannot be allowed in decoding new sentences⇒ a hard decision must be taken regarding tuple segmentation



- Baseline criterion: IBM model 1 score for each possible tuple
- New criterion: entropy of Part-Of-Speech distributions

Linguistic tuple segmentation

Forward entropy

Probability of observing a certain Part-Of-Speech following the sequence of words defined by t_{i-1} and t_i:

$$p^f_{POS} = \frac{N(t_{i-1}, t_i, POS_{i+1})}{\sum_{POS'} N(t_{i-1}, t_i, POS'_{i+1})}$$

Entropy of the POS distribution in position $i+1$ given (t_{i-1}, t_i):

$$H^f_{POS} = -\sum_{POS} p^f_{POS} \log p^f_{POS}$$

Backward entropy

Similarly, calculate a "backward" entropy of POS distribution preceding (t_i, t_{i+1}).

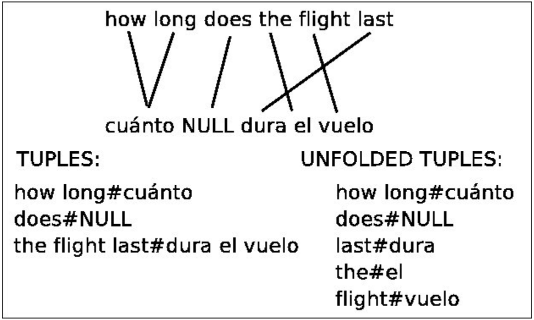
Linguistic tuple segmentation

if $H^f_{POS} > H^b_{POS}$, we have observed (t_{i-1}, t_i) in more grammatically different contexts than (t_i, t_{i+1}).

$\Rightarrow$ t_{i-1} and t_i tend to be more often connected than t_i and t_{i+1}, and should belong to the same translation tuple.

Tuple unfolding

Before reordering search, extract tuples with an unfolding technique



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

- Gives smaller tuples, thus easier to re-use
- Gives higher probability to bilingual n-grams with correct target language order

Constrained reordered search

Basic reordered search exploring all possibilities, with restrictions:

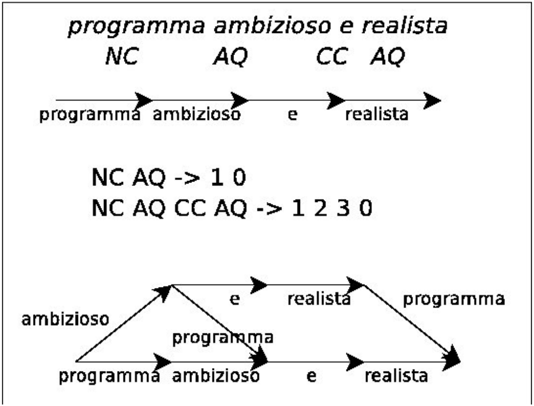
- Distortion limit (m): Any tuple is only allowed to be reordered within a limited distance (in number of source words).
- Reordering limit (j): Any translation path is only allowed to perform j reordering jumps.

for IWSLT 2006, given the average sentence length, we set $m = 5$ and $j = 3$ for all language pairs

When this word ordering strategy was applied, a simple word distance-based distortion model was added as an additional feature to the system.

Reordering patterns

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



Pattern extraction

Pattern instances are automatically learnt in training from the crossed links found in tuples (in a way equivalent to unfolding)

Decision to prune out or use each pattern based on relative frequency:

$$p(t_1, ..., t_n \mapsto i_1, ..., i_n) = \frac{N(t_1, ..., t_n \mapsto i_1, ..., i_n)}{N(t_1, ..., t_n)}$$

(this probability is not used in decoding. Only in training, to prune out some patterns)

Experiments description

- alignment: IBM model 4 union (GIZA++ [Och, 2000]), 50 classes (mkcls), lowercased
- bilingual and target language models: standard 4-gram models (SRILM [Stolcke, 2002])
- preprocessing: split sentences at dots (if equal number of dots)
- language-dependent preprocessing: see paper

Results

official	test		ASRtest	
	BLEU	NIST	BLEU	NIST
Arabic→English				
P: m5j3	**0.232**	**6.24**	**0.214**	**5.82**
C1: rgraph	0.227	6.14	0.205	5.69
C2: m5j3 segIBM	0.227	6.06	0.210	5.63
C3: m5j3 lm20	0.225	6.13	0.205	5.71

Results

official	test		ASRtest	
	BLEU	NIST	BLEU	NIST
Arabic→English				
P: m5j3	**0.232**	**6.24**	**0.214**	**5.82**
C1: rgraph	0.227	6.14	0.205	5.69
C2: m5j3 segIBM	0.227	6.06	0.210	5.63
C3: m5j3 lm20	0.225	6.13	0.205	5.71
Italian→English				
P: rgraph alem	**0.333**	**7.75**	**0.282**	**6.87**
C1: rgraph	0.331	7.63	0.278	6.75
C2: rgraph segIBM	0.332	7.64	0.273	6.71
C3: rgraph lm20	0.323	7.54	0.271	6.69

Results

official	test		ASRtest	
	BLEU	NIST	BLEU	NIST
Arabic→English				
P: m5j3	**0.232**	**6.24**	**0.214**	**5.82**
Italian→English				
P: rgraph alem	**0.333**	**7.75**	**0.282**	**6.87**
Chinese→English				
P: m5j3	**0.186**	5.57	**0.162**	4.98
C1: rgraph	0.183	**5.74**	0.157	**5.12**

Results

official	test BLEU	test NIST	ASRtest BLEU	ASRtest NIST
Arabic→English				
P: m5j3	**0.232**	**6.24**	**0.214**	**5.82**
Italian→English				
P: rgraph alem	**0.333**	**7.75**	**0.282**	**6.87**
Chinese→English				
P: m5j3	**0.186**	5.57	**0.162**	4.98
C1: rgraph	0.183	**5.74**	0.157	**5.12**
Japanese→English				
P: rgraph	0.146	**5.27**	0.137	**4.94**
C1: m5j3	**0.152**	5.18	**0.141**	4.89

Conclusions and further work

Basically two novel features were introduced in our system.

- Extension of monotonic search graph with reordered paths suggested by POS-tags-based patterns:
  - dramatic efficiency improvement (nearly as efficient as monotonic search)
  - outperforms constrained reordered search for Italian→English, achieves similar results for Chinese→English and Japanese→English and is slightly worse in Arabic→English
  - thus, these patterns don't capture long reordering (in this case, POS-tag-based patterns lead to sparseness problems)
  - further work should focus on pattern extraction for language pairs demanding long reorderings (e.g. syntax-based patterns)
- tuple segmentation based on POS entropy: yields a slight yet systematic improvement in translation quality

Other direction for further research: better integration of speech recognition output (word lattices, N-best lists)