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:

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

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

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

how long does the flight last

V / X

cuánto NULL dura el vuelo

TUPLES:
how long#cuánto
does#NULL
the flight last#dura el vuelo

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:

- 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
1. TALP Ngram-based Translation System

2. Tuple segmentation strategies
   - Introduction
   - Linguistic Tuple Segmentation

3. Word ordering strategies

4. Experiments

5. Conclusions and Further Work
**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 $\Rightarrow$ 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 $t_i$:

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

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

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

**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.
Word ordering strategies

1. TALP Ngram-based Translation System

2. Tuple segmentation strategies

3. Word ordering strategies
   - Tuple unfolding
   - Constrained reordered search
   - Reordering Patterns

4. Experiments

5. Conclusions and Further Work
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

```
programma ambizioso e realista
NC AQ CC AQ
NC AQ -> 1 0
NC AQ CC AQ -> 1 2 3 0
```
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:

\[ p(t_1, \ldots, t_n \mapsto i_1, \ldots, i_n) = \frac{N(t_1, \ldots, t_n \mapsto i_1, \ldots, i_n)}{N(t_1, \ldots, t_n)} \]

(this probability is not used in decoding. Only in training, to prune out some patterns)
1. TALP Ngram-based Translation System

2. Tuple segmentation strategies

3. Word ordering strategies

4. Experiments
   - Description
   - Results

5. Conclusions and Further Work
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

<table>
<thead>
<tr>
<th>Arabic→English</th>
<th>official</th>
<th>test</th>
<th>ASRtest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>P: m5j3</td>
<td>0.232</td>
<td>6.24</td>
<td>0.214</td>
</tr>
<tr>
<td>C1: rgraph</td>
<td>0.227</td>
<td>6.14</td>
<td>0.205</td>
</tr>
<tr>
<td>C2: m5j3 segIBM</td>
<td>0.227</td>
<td>6.06</td>
<td>0.210</td>
</tr>
<tr>
<td>C3: m5j3 lm20</td>
<td>0.225</td>
<td>6.13</td>
<td>0.205</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>official</th>
<th>test</th>
<th>ASRtest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic→English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: m5j3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1: rgraph</td>
<td>0.227</td>
<td>6.14</td>
<td>0.205</td>
</tr>
<tr>
<td>C2: m5j3 segIBM</td>
<td>0.227</td>
<td>6.06</td>
<td>0.210</td>
</tr>
<tr>
<td>C3: m5j3 lm20</td>
<td>0.225</td>
<td>6.13</td>
<td>0.205</td>
</tr>
<tr>
<td>Italian→English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: rgraph alem</td>
<td>0.331</td>
<td>7.63</td>
<td>0.278</td>
</tr>
<tr>
<td>C1: rgraph</td>
<td>0.331</td>
<td>7.63</td>
<td>0.278</td>
</tr>
<tr>
<td>C2: rgraph segIBM</td>
<td>0.332</td>
<td>7.64</td>
<td>0.273</td>
</tr>
<tr>
<td>C3: rgraph lm20</td>
<td>0.323</td>
<td>7.54</td>
<td>0.271</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Official</th>
<th>Test</th>
<th>ASR Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
</tr>
<tr>
<td>Arabic→English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: m5j3</td>
<td>0.232</td>
<td>6.24</td>
<td>0.214</td>
</tr>
<tr>
<td>Italian→English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: rgraph alem</td>
<td>0.333</td>
<td>7.75</td>
<td>0.282</td>
</tr>
<tr>
<td>Chinese→English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: m5j3</td>
<td>0.186</td>
<td>5.57</td>
<td>0.162</td>
</tr>
<tr>
<td>C1: rgraph</td>
<td>0.183</td>
<td>5.74</td>
<td>0.157</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Official</th>
<th>Test</th>
<th>ASRtest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arabic → English</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: m5j3</td>
<td>0.232</td>
<td>6.24</td>
</tr>
<tr>
<td><strong>Italian → English</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: rgraph alem</td>
<td><strong>0.333</strong></td>
<td><strong>7.75</strong></td>
</tr>
<tr>
<td><strong>Chinese → English</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: m5j3</td>
<td><strong>0.186</strong></td>
<td>5.57</td>
</tr>
<tr>
<td>C1: rgraph</td>
<td>0.183</td>
<td><strong>5.74</strong></td>
</tr>
<tr>
<td><strong>Japanese → English</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: rgraph</td>
<td>0.146</td>
<td><strong>5.27</strong></td>
</tr>
<tr>
<td>C1: m5j3</td>
<td><strong>0.152</strong></td>
<td>5.18</td>
</tr>
</tbody>
</table>
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)