Reordering via N-Best Lists for Spanish-Basque Translation

Germán Sanchis and Francisco Casacuberta
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia, Spain

Introduction

• SMT systems are an efficient way of building cheap and good quality machine translation systems.

Statistically, the machine translation problem can be defined as

\[ \text{score}(s|t) = \text{argmax} \{ \text{Pr}(s|t) \} \]

• The previous probability can be decomposed into the target statistical language model and the translation model

\[ \text{score}(s|t) = \text{argmax} \{ \text{Pr}(s) \cdot \text{Pr}(t|s) \} \]

• The model is applied by reordering the source sentence, \( s \), to give a target sentence, \( t \).

• State of the art SMT systems rely on phrase based models.

• Such systems often incur in word ordering related errors.

• Ordering errors lead to incorrect translations, but also incorrectly estimated parameters.

• Distortion models usually implemented within the decoding algorithm, completing computational problems and ultimately restrictions are being applied. Hence search turns sub-optimal and representational power is lost.

• Arbitrary word reorderings lead to NP-hard search.

Brief overview of existing approaches

• Output sentence reordering: two main approaches:
  - J.M. Vilà et al.
    1. Monotonic most probable non-monotone alignment patterns and add a mark to “remember” original order
    2. Train new system with this setup and reorder output according to the marks found in the output sentence.
  - Kumar and Byrne learned WFEs accounting for local reorderings of two or three phrase positions. Training the model did not yield statistically significant results w.r.t. the introduction of the models with fixed probabilities.

• Input sentence reordering:
  - Developed at the RWTH-Aachen
  - Idea: avoid the non-monotonic translation problem by reordering the input sentence.
  - Alignments made to establish which word order is appropriate for monosyllabic translation.
  - In search not possible, testing all permutations prohibitive: local IBM, inverse IBM and ITG restrictions.
  - However, search space still huge, and a very high computational power is paid.

The reordering model and N-Best reorderings

• The method described above disregards the information contained within monotonous corpora, which can be used to train a reordering model with the aim of reordering monotonous corpora.

• Corpus monotonization: Algorithm

  1. Let \( s \) be a source sentence, and \( s_j \) its j-th word.
  2. Let \( t \) be a target sentence, and \( t_k \) its k-th word.
  3. Let \( C \) be a cost matrix \( C_{ij} = c(s_j, t_k) \).
  4. Let \( \{ s' \} = \{ \text{all possible permutations of } s \} \).
  5. Compute alignment \( A_g(j) = \text{argmin}_{i} C_{ij} \).
  6. \( s' = \{ s' \} \cdot \{ A_g(j) \} \cdot \{ s' \} + 1 \).
  7. Re-compute \( C \), obtaining \( C' \).
  8. \( \text{optimal} \) \( C' \) = argmin_{ij} C'_{ij}.

• Monotonized alignments define a new source language.

  \( \rightarrow \) A reordered language model can be trained with the reordered input sentences \( s' \).

  \( \rightarrow \) Same vocabulary as the source language.

  \( \rightarrow \) Same word order as the target language.

  \( \rightarrow \) Reordering model will most likely not depend on the output sentence.

• Example:

  10em
dia
ei
noso

10em
dia
ei
noso

Reordering problem can be defined as follows:

\[ s' = \text{argmax} \{ \text{Pr}(s'|t) \cdot \text{Pr}(s'\mid s) \} \]

where \( \text{Pr}(s'|t) \) is the reordered language model, and \( \text{Pr}(s'|s) \) is the reordering model.

Reordering problem very similar to the translation problem, but with a very constrained translation table.

\( \rightarrow \) Same methods developed to solve the translation problem can be used to face the reordering problem.

Exponential reordering model, defined as:

\[ \text{Pr}(s|t) = \exp(-\sum_{i,j} C_{ij}) \]

To reduce the error that the reordering model introduces, we compute an n-best list of reordering hypothesis and translate them all, selecting as final output the one which the best score according to \( \text{Pr}(s|t) \).

Ultimately, we are constraining the search space of permutations of the source sentence much more than previous approaches, while taking into account the information that monotonized alignments entail.

Experimental setup

• First, the bilingual pairs were aligned using IBM model 4 by means of the GIZA++.

• Alignments made monotone as described, new alignment determining the new monotonous alignment is calculated.

• Reordered source sentence model built; phrase extraction is performed by means of the Thot toolkit.

• Pharaoh toolkit used as reordering model.

• Translation table only includes the words contained in the vocabulary of source language.

• Toolkit reorders the words by taking into account the language model and exponential model.

• Since the phrases are just words, result is an exponential word-reordering model.

• The best reordering hypothesis are translated using Pharaoh.

• Best scoring one is kept, where the score is the product of the (inverse) translation model and the language model.

• As baseline, the same pipeline without reordering steps: GIZA++ for aligning, Thot for phrase extraction and Pharaoh for translating.

Translation results

The system described was tested on a Basque-Spanish translation task, more specifically the Tourist corpus:

<table>
<thead>
<tr>
<th>Corpus characteristics</th>
<th>Spanish</th>
<th>Basque</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>38940</td>
<td>20138</td>
</tr>
<tr>
<td>Different pairs</td>
<td>16759</td>
<td>10516</td>
</tr>
<tr>
<td>Words</td>
<td>16614</td>
<td>9800</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>772</td>
<td>850</td>
</tr>
<tr>
<td>Average length</td>
<td>9.5</td>
<td>9.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
</tr>
<tr>
<td>BLEU</td>
</tr>
<tr>
<td>WER</td>
</tr>
<tr>
<td>FER</td>
</tr>
</tbody>
</table>

PER criterion improvement prove that better phrases are extracted due to the input sentence reordering.

• Translation quality from Spanish to Basque much higher than vice-versa because of corpus characteristics.

• Increasing n yields better results for Spanish-Basque.

• Only using the best hypothesis already better results than baseline.

Conclusions

• A reordering technique has been implemented, taking profit of the information in monotonous corpora.

• By reordering, better quality phrases can be extracted, improving performance for languages with heavy reordering.

• This technique has been applied to translate a semi-synthetic Spanish-Basque corpus, with promising results.

• The technique proposed is learnt automatically, without linguistic annotation or manual reordering rules.

• Both reordering corpora and reordering techniques seem to have a very important potential.

Future work

• Obtaining results with non-synthetic and richer corpora.

• Perform experiments on other language pairs, involving Arabic, Japanese or Chinese.

• Development of more specific reordering models, more suitable for this task.

• Investigating integrated approaches for solving the reordering problem.

Acknowledgements

This work has been partially supported by the EC (FREDER) and the Spanish MEC under grant TIN2006-15694-C02-01 and by the MIC scholarship AP2005-4923.