Reordering via N-Best Lists for Spanish-Basque Translation

Germán Sanchis and Francisco Casacuberta

Instituto Tecnológico de Informática
Universidad Politecnica 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:

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, Pr(t|s)$$

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

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, Pr(t) \cdot Pr(s|t)$$

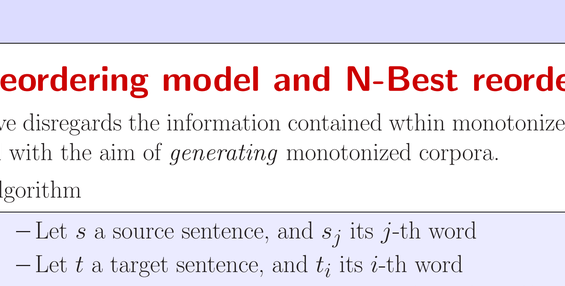
- State of the art SMT systems rely on phrase based models.
- Such systems often incur in word ordering related errors.
- Ordering errors lead to incorrect translations, but also incorrectly estimated parameters.
- Distortion models usually implemented within the decoding algorithm, implying computational problems and ultimately restrictions being applied. Hence search turns sub-optimal and representational power is lost.
- Arbitrary word reoderings lead to NP-hard search.

Brief overview of existing approaches

- Output sentence reordering: two main approaches.
  - J.M.Vilar et al.:
    1. Monotonize 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 WFSTs accounting for local reorderings of two or three phrase positions. Training the models 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-monotonous translation problem by reordering the input sentence
  - Alignment models used to establish which word order is appropriate for monotonous 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 price is paid.

The reordering model and N-Best reorderings

- The method described above disregards the information contained wthin monotonized corpora, which can be used to train a reordering model with the aim of *generating* monotonized corpora.
- Corpus monotonization: Algorithm
  - Let s a source sentence, and s_j its j-th word
  - Let t a target sentence, and t_i its i-th word
  - Let C be a cost matrix s.t. $c_{ij} = \text{cost}(\text{align}(s_j, t_i))$
  - Let $\{s^r\} = \{\text{all possible permutations of } s\}$.
    1. compute alignment $A_D(j) = \underset{i}{\operatorname{argmin}}\, c_{ij}$
    2. $s' = \{s^r | \forall j : A_D(j) \leq A_D(j+1)\}$
    3. recompute (reorder) C, obtaining C'.
    4. set $A'_I(i) = \underset{j}{\operatorname{argmin}}\, c'_{ij}$.
    5. Optional: Compute minimum-cost *monotonic* path through cost matrix C'.
- Monotonized alignments define a new source *language*.
  - → A reordered language model can be trained with the reordered input sentences s'.
  - → Same vocabulary as the source language.
  - → Same word order as the target language.
  - → Reordering model will most likely not depend on the output sentence.
- Example:



- Reordering problem can be defined as follows:

$$s' = \underset{s^r}{\operatorname{argmax}}\, Pr(s^r) \cdot Pr(s|s^r)$$

  where $Pr(s^r)$ is the reordered language model, and $Pr(s|s^r)$ is the reordering model.
- Reordering problem very similar to the translation problem, but with a very constrained translation table.
  - → Same methods developed to solve the translation problem can be used to face the reordering problem.
  - → Exponential reordering model, defined as:

$$Pr(s|s') \approx exp(-\sum_i d_i)$$

  - → 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 whith the best score according to $Pr(t) \cdot Pr(s^r|t)$.
- 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 exponenetial word-reordering model.
- n 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 Pharoah for translating.

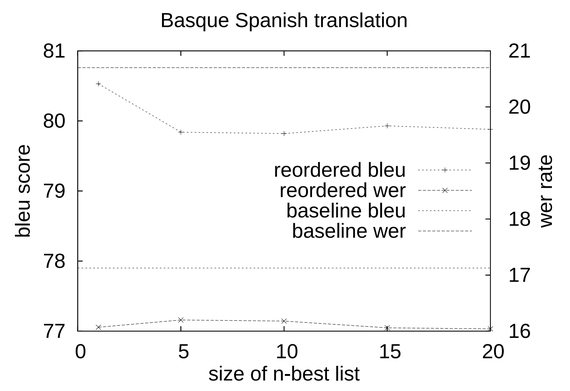
Translation results

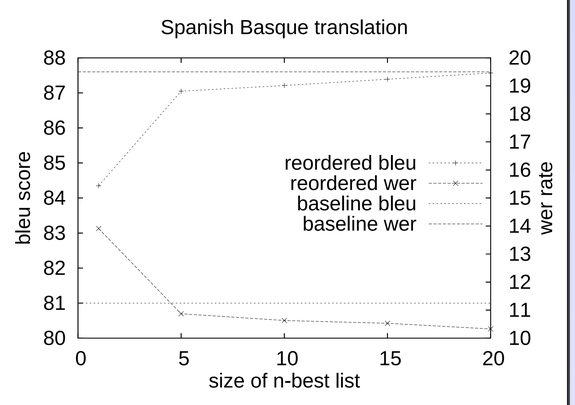
- The system described was tested on a Basque-Spanish translation task, more specifically the Tourist corpus:

Corpus characteristics:

		Spanish	Basque
Training	Sentences	38940	
	Different pairs	20318	
	Words	368314	290868
	Vocabulary	722	884
	Average length	9.5	7.5
Test	Sentences	1000	
	Test independent	434	
	Words	9507	7453
	Average length	9.5	7.5

- Results:





	Baseline	Reordered, $n = 5$
WER	20.7%	16.2%
BLEU	77.9%	79.8%
PER	12.6%	11.0%

	Baseline	Reordered, $n = 5$
WER	19.5%	10.9%
BLEU	81.0%	87.1%
PER	6.2%	4.9%

- PER criterion improvement proves 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 monotonized 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 promissing results.
- The technique proposed is learnt authomatically, 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 (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01 and by the MEC scholarship AP2005-4023.