LIMSI@WMT13

Alexandre Allauzen, Nicolas Pécheux, Quoc Khanh Do, Marco Dinarelli, Thomas Lavergne, Aurélien Max, Hai-Son Le, François Yvon Univ. Paris Sud and LIMSI—CNRS Orsay, France

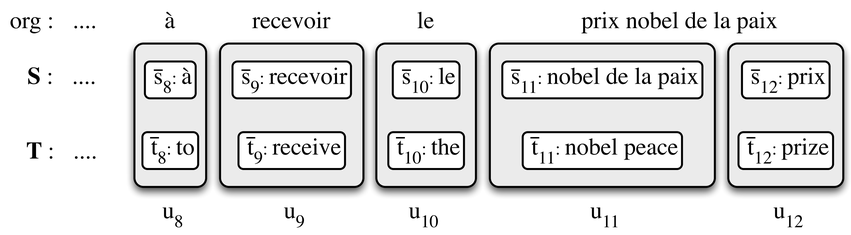
Highlights

{French,German,Spanish}↔English shared translation tasks

- *n*-code (http://ncode.limsi.fr):
  - Source reordering and *n*-gram translation models (TMs)
  - *n*-best reranking with SOUL LM and TM
- First participation to the Spanish-English task
- Preliminary experiments with source pre-ordering
- Tighter integration of continuous space language models
- Different tuning sets according to the original language

n-code

Tuples are bilingual units



The translation model is a *n*-gram model of tuples

n-code models

- 3-gram tuple LM and 4-gram target word LM
- Four lexicon models (similar to the phrase table)
- Two lexicalized reordering models (orientation of next/previous translation unit)
- Weak distance-based distortion model
- Word-bonus and a tuple-bonus

Data Processing

- Better normalization tools provide better BLEU scores
- Specific pre-processing for German as source language
- Language model tuning
- Using lemmas and POS for the Spanish-English (FREELING[1])
- Cleaning noisy data sets (*CommonCrawl*)
  - Filter out sentences in other languages
  - Sentence alignment
  - Sentence pair selection using a perplexity criterion

[1]: http://nlp.lsi.upc.edu/freeling/

Artificial Text generation with SOUL

Issue:

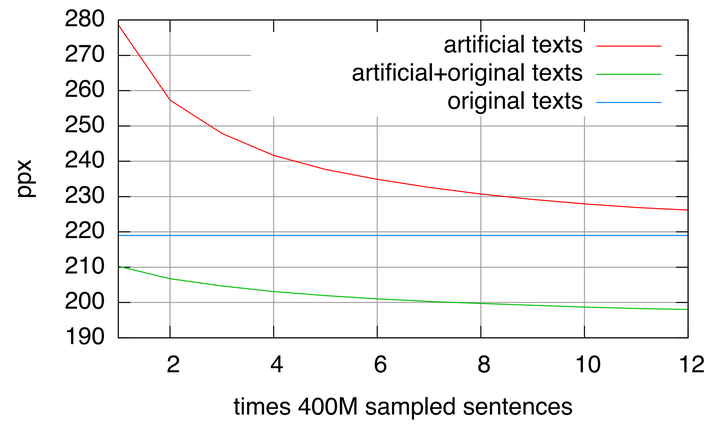
- Conventional models are used to build the pruned search space and the potentially sub-optimal k-best lists
- Continuous space models can only be used to rerank k-best lists

A solution for the language model:

- Sample text from a 10-gram NNLM
- Estimate a conventional 4-gram model used by the decoder.

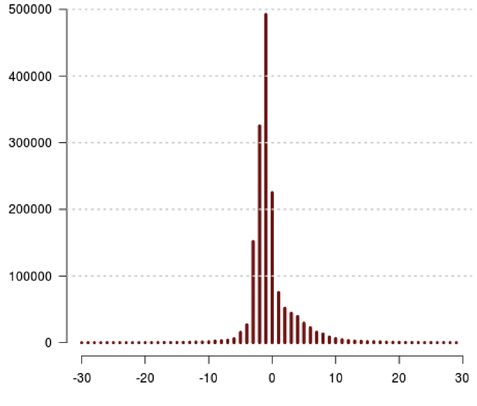
Experiments:

- On the English to German task



Pre-ordering (English to German)

- ITG-like parser[1] to generate reordered source sentences
- Build a SMT system using pre-ordered parallel data
- Only 16% of the source training sentences are modified
- No clear impact



Histogram of token movement size

[1] :http://www.phontron.com/lader/

Tuning *vs* original language

Assumption:

- The original language of a text has a significant impact on translation performance

Solution:

- Different tunings for different original languages:

Original language	Baseline *fr2en*	Tuning per source
cz	22.3	23.8
en	36.4	39.2
fr	31.6	32.4
de	18.5	18.5
es	30.2	29.3
all	29.4	30.1

Experimental Results

Direction	System	BLEU	
		dev nt09	*test nt10*
en2de	base	15.3	16.5
	+artificial text	15.5	16.8
	+SOUL	16.4	17.6
	+artificial text+SOUL	16.5	17.8
de2en	base+SOUL	22.8	24.7
en2fr	base+SOUL	29.3	32.6
fr2en	base	29.1	29.4
	tuning *per* source		30.1
	+SOUL	30.1	30.6
		dev nt11	test nt12
en2es	base+pos	32.3	33.8
es2en	base+pos+lem	30.7	33.9