LIMSI @WMT11

A. Allauzen, G. Adda, H. Bonneau-Maynard, J.M Crego, H-S Le,
A. Sokolov, A. Max, A. Lardilleux, G. Wisniewski, T. Lavergne, F. Yvon
Univ. Paris Sud and LIMSI—CNRS Orsay, France

Highlights

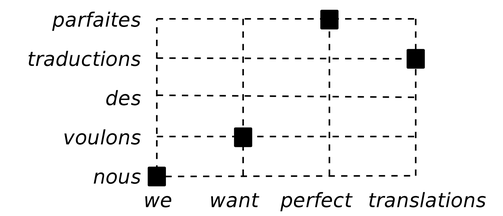
French-English and German-English shared translation tasks in both directions

- n-code: open source Statistical Machine Translation system
  - Source reordering as pre-processing
  - Translation model based on bilingual n-grams
- **Simple and efficient filtering strategy** of the *GigaWord*.
- **First use of the SOUL** target language model in SMT

  ⇒ significant improvements with 10-gram models

n-CODE

Tuples are bilingual units



(1) we nous	want voulons	NULL des	translations traductions	perfect parfaites
(2) we nous	want voulons	translations des_traductions		perfect parfaites

n-code's model

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

Data Pre-processing

- Better normalization tools provide better BLEU scores
- Specific pre-processing for German as source language
- Cleaning noisy data sets (*GigaWord*)
  - Discard sentences in other languages
  - Remove repeated sentences, or the ones included in the development sets

    → for the monolingual news data, this can reduce the amount of data by a factor 3 or 4
  - Normalize the character set

Data filtering

Filtering the *GigaWord* Corpus

For each side:

- Train a specific language model on a selection of news texts
- Rank sentences according to their perplexity
- Select sentences above a given threshold

Two thresholds:

- The upper quartile ⇒ a subset of 25% (6.7M of sentences)
- The median ⇒ a subset of 50%

Baseline Results (*newstest2010*)

- **Filtering the *GigaWord* corpus** for French-English:

System	en2fr		fr2en	
	BLEU	*TER*	BLEU	*TER*
All	27.4	56.6	26.8	55.0
Upper quartile	27.8	56.3	28.4	53.8
Median	**28.1**	**56.0**	**28.6**	**53.5**

- German-English, **impact of the the POS tagger**:

	en2de		de2en	
	BLEU	*TER*	BLEU	*TER*
RFTagger	22.8	60.1	**16.3**	66.0
TreeTagger	**23.1**	59.4	16.2	66.0

Target n-gram Language Model

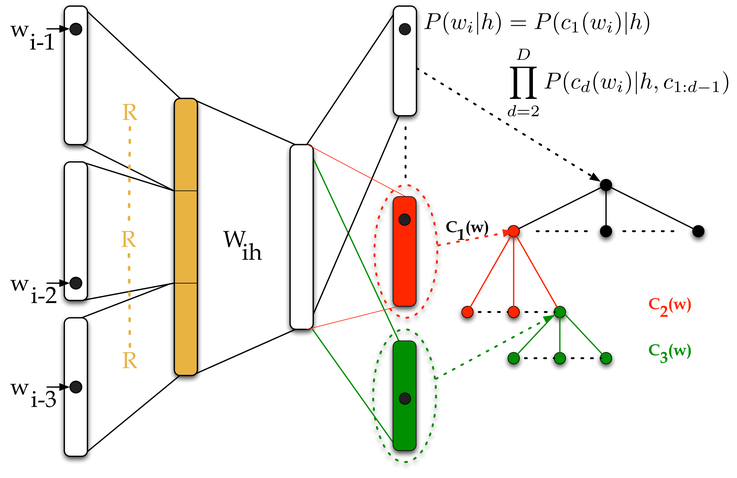
Standard 4-gram Back-off Language Models

- Total of running words: 2.5G in French and 6.2G in English
- Using all the available data.
- Linear interpolation of several LMs
- Specifically tuned for news text of 2010

The soul LM

- A large vocabulary continuous space LM.
- Use a clustering tree to structure the output vocabulary.
- The order n can be raised without a prohibitive increase in complexity.

SOUL Language Model Overview



Results with various LMs

- Linear interpolation of 4 SOUL LMs (different re-sampling)
- Initial shortlist of $5k$ words
- K-means recursive word clustering based on the continuous representation of words (R), depth of tree = 3
- n-best rescoring, tuned on *newstest2009*

SOUL LM	en2fr		en2de	
	BLEU	*TER*	BLEU	*TER*
without	28.1	56.0	16.3	66.0
4-gram	28.4	55.5	16.5	64.9
6-gram	28.7	55.3	16.7	64.9
10-**gram**	**28.8**	**55.2**	**16.8**	**64.6**