LIMSI @ IWSLT 2010

Alexandre Allauzen, Josep M. Crego, İlknur Durgar El-Kahlout,
Le Hai-Son, Guillaume Wisniewski and François Yvon

N-Code highlights

A billingual n-gram based decoder

- Each training sentence pair is a unique sequence of tuples with a minimal segmentation.
- Source side reordering computed before decoding via POS-based rewrite rules

Abstract

The Talk task (English to French):

- Extension of our in-house N-Code SMT system: a bilingual reordering model over *generalized translation units*
- Use of training data extracted from Wikipedia for target language model adaptation

The BTEC task (Turkish to English):

- *Moses* based system
- Pre-processing schemes for Turkish to reduce the morphological discrepancies with English
- Continuous space language models

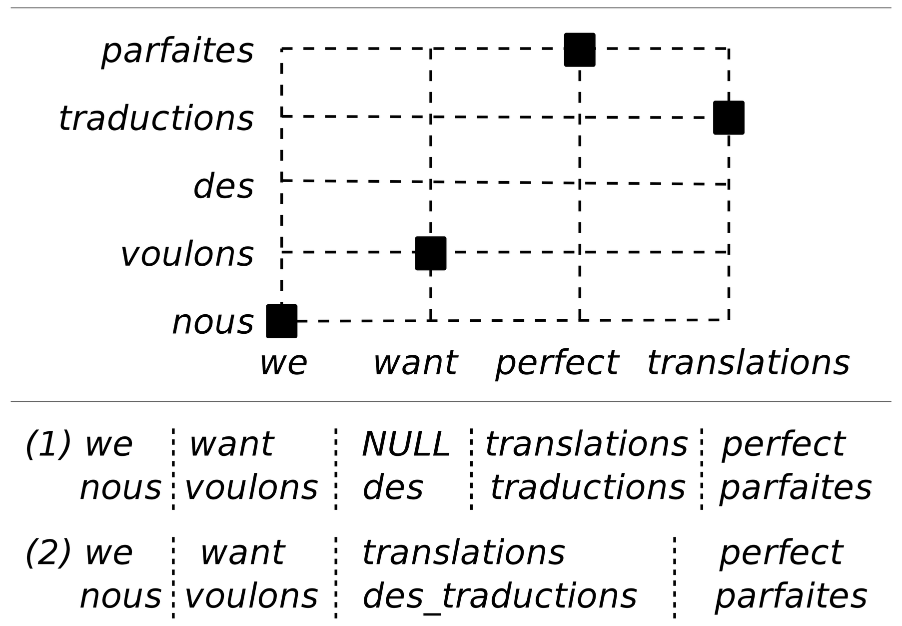
Turkish specificities

- The productive and agglutinative morphology of Turkish implies a large vocabulary size
- Turkish has a flexible word order, but mainly subject-object-verb (SOV).

⇒ may affect the reliability of word alignment and the phrase extraction.

System description

Tuples are bilingual units



N-Code's model

- Tuple 3-gram and target word 4-gram LMs (Kneser-Ney & interpolation)
- Two lexicon models (complementary translation scores for each tuple)
- Two lexicalized reordering models (predict orientation of next/previous translation unit)
- A weak distance-based distortion model
- A word-bonus and a tuple-bonus models

Basic pre-processing

- **Tokenization**: For both languages, we used in-house tokenizers
- Turkish texts are morphologically analyzed and disambiguated
- Turkish words are represented with stems and lexical morphemes and then lowercased:

  evin (your house) → *ev+in*

- English texts are in true-case.

Frequency based segmentation

A recursive morphological decomposition:

ver+ma+dh+m (I didn't give) → *ver+ma+dh* *+m*

- the whole word is segmented if the frequency of *ver+ma+dh* is upon a given threshold
- then we similarly consider the split and so on:

  ver+ma+dh → *ver+ma* *+dh*

The best BLEU improvement was achieved with a threshold of 10.

A bilingual n-gram reordering model

we	want	translations	perfect
nous	voulons	des_traductions	parfaites

pronoun	verb	noun	adjective
pronoun	verb	det_noun	adjective

- n-gram LM over generalized translation units
- Generalized translation units enables larger n-gram contexts (up to 6-grams)
- Helps to capture mid-range syntactic reorderings
- "Translation model" sentence structure

Out-of-vocabulary words

Before tuning and decoding:

- Similar to the segmentation, OOV words are split morpheme by morpheme to get a "known" word
- When the root word is OOV, the whole word with all its morphemes is removed
- Exception for proper nouns

Other pre-processing steps

- Question inversion
- Short distance morpheme reordering
- Augmented training data with open-class words
  - bias content root word alignments
  - consider open-class words in both sides

Experimental results

Configuration	talk-tune	talk-test
base	35.82	35.35
base+bil6g	35.60	35.22
base-wikipedia	35.40	35.32

Wikipedia target LM

- The data extracted from the French Wikipedia is roughly extracted, filtered and tokenized
- A total amount of 40M tokens
- A specific target LM is estimated

- The interpolation weights for LM are tuned on a held-out subset of the training *TED-1.1 corpus*
- The official TALK dev. is divided in two parts (*talk-tune* and *talk-test*)
- The system is tuned with MERT on *talk-tune*
- All configurations achieve very similar results
- The bilingual reordering model, as well as the use of Wikipedia do not yield to a significative BLEU improvement.

Experimental results

Impact of the pre-processing schemes:

System	BLEU
Baseline−3g-lm	37.15
Baseline−4g-lm	37.21
Baseline−5g-lm	**38.37**
Segmentation−t10	50.06
+Question Inversion	48.85
+Local Ordering	49.74
+Content Words	51.72
+OOV	57.25

Final System:

- Augmented training data using multiple references
- Frequency based segmentation, open-class words, splitting of OOV words
- 7-gram standard neural-network language model in a two pass decoding approach

System	case+punc. BLEU
iwslt09	52.97
iwslt10	48.42