The TALP&I2R SMT Systems for IWSLT 2008

Maxim Khalilov, Marta R. Costa-jussà, Carlos A. Henríquez, José A.R. Fonollosa, Adolfo Hernández, José B. Mariño, Rafael E. Banchs, Chen Boxing, Min Zhang, Aiti Aw and Haizhou Li

1TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
2Department of Human Language Technology
Institute for Infocomm Research, Singapore

ABSTRACT

UPC TALP Research Center participated in the Arabic-English task and together with the I2R participated in Chinese-Spanish translation and pivot Chinese-(English)-Spanish translation. The novelties we have introduced are:

1. improved reordering method for an Ngram-based system,
2. linear combination of translation, reordering and target models for domain adaptation,
3. new technique dealing with punctuation marks insertion, and
4. concatenation strategy for PIVOT translation for a phrase-based SMT system.

1 BASELINE SYSTEMS

- Bilingual Ngram Translation Model (Marino et al. CL’06) (TALPpuples)
  - The translation model is based on bilingual n-grams.
  - Bilingual units, i.e. tuples, are extracted from a word-to-word aligned corpus according to:
  1. Tuple extraction should produce a monotonic segmentation of bilingual sentence pairs;
  2. No smaller tuples can be extracted without violating the previous constraint.

- Bilingual Phrase Translation Model: MOSES System [Koehn et al. 07] (TALPphrases)
  - The translation model is based on phrases.
  - Bilingual units, i.e. phrases, are extracted from a word-to-word aligned corpus according to:
  1. Words are consecutive along both sides of the bilingual phrase.
  2. No word on either side of the phrase is aligned to a word out of the phrase.

- Feature functions: in addition to the translation model, the baseline system implements a combination of feature functions.

2 REORDERING TECHNIQUE (SMR)

- The conception of the Statistical Machine Reordering (SMR) stems from the idea of using the powerful techniques developed for SMT and to translate the source language (S) into a reordered source language (‘S’), which more closely matches the order of the target language.

- To infer more reorderings, it makes use of word classes and to correctly integrate the SMT and SMR systems, both are concatenated by using a word graph which offers weighted reordering hypotheses to the SMT system.

3 ARABIC-TO-ENGLISH TASK

3.1 TRANSLATION INTERPRETATION (POST-EVALUATION)

- We used an out-of-domain corpus to increase the final translation and reordering results. We performed a linear combination of the translation, reordering and target models.

3.2 PUNCTUATION RESTORATION (PRIMARY)

- We embedded punctuation restoration in the main translation step.

3.3 EXPERIMENTS

- MADA-TOKAN system for disambiguation and tokenization.
- The out-of-domain was a 136K-line subset from the Arabic News, English Translation of Arabic Treebank and Ummah LDC parallel corpora (VIOLIN) [Habash et al. 08]
- Primary system: the TALPphrases MOSES-based system enhanced with the punctuation marks repetition technique.
- Secondary system: TALPpuples system, configured to use the bilingual TM of order 4, 4-gram target-side LM and 4-gram POS target-side LM. It includes SMT with 100 statistical classes.
- Post-translation system: the TALPphrases MOSES-based system enhanced with the punctuation marks repetition and interpolation technique.

4 CHINESE-(ENGLISH)-SPANISH PIVOT TRANSLATION

4.1 SYSTEM CASCADE (PRIMARY)

- Using the 50-best list of translation hypotheses generated by the decoder for the Chinese-to-English system.
- A 4-best list was made for each of the first list instances, totally representing a 200-best of possible Spanish translations for each Chinese sentence.
  The single-best translation was computed using a Minimum Bayes Risk (MBR) strategy [Kumar et al. 04]

4.2 PHRASE PROBABILITIES COMBINATION (SECONDARY)

- Combination of the phrase translation probabilities of the two language pairs (Chinese-English and English-Spanish translations) with the strategy proposed [Wu and Wang, 2007] to obtain the translation probabilities for each Chinese-Spanish phrase. The final phrase probabilities were calculated as follows:

\[
\text{\( P(j|e) = \sum_i P(j|i)\cdot P(e|i) \)}
\]

4.3 EXPERIMENTS

- Word segmentation for the Chinese part using ICTCLAS tools.
- For the Chinese-English, the out-of-domain corpora was: the HIT corpus (132K sentence pairs); Olympic corpus (54K bilingual sentences); PKU-corpus (200K parallel phrases) and the English part of the Tanaka corpus.

5 CHINESE-TO-SPANISH DIRECT TRANSLATION

5.1 EXPERIMENTS

- Primary system: TALPpuples system, configured as in the Arabic-English task.
- Secondary system: the TALPphrases MOSES-based system.

6 CONCLUSIONS

- Arabic-English: the domain adaptation using linear interpolation of translation, reordering and target models shows improvements in CRR and ASR.
- Chinese-(English)-Spanish: the system cascade architecture demonstrates better results than the alternative model (phrase probabilities combination), however there is still room for improvement on phrase table pruning.
- Chinese-Spanish: Although the direct Chinese-Spanish phrase-based system operated better than the TALPpuple system on the internal test, we submitted the last one as a primary system in order to contrast it the many other MOSES-based strategies presented in the evaluation.