The TALP&I2R SMT Systems for IWSLT 2008

Maxim Khalilov[1], *Marta R. Costa-jussà*[1], *Carlos A. Henríquez*[1], *José A.R. Fonollosa*[1],*Adolfo Hernández*[1], *José B. Mariño*[1], *Rafael E. Banchs*[1],*Chen Boxing*[2], *Min Zhang*[2], *Aiti Aw*[2] *and Haizhou Li*[2]

[1]TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
{khalilov|mruiz|carloshq|adrian|adolfohh|canton|rbanchs}@talp.upc.edu

[2]Department of Human Language Technology
Institute for Infocomm Research, Singapore
{bxchen|mzhang|aaiti|hli}@i2r.a-star.edu.sg

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.

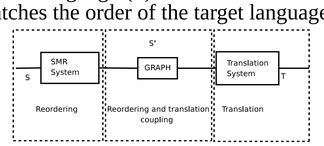
1 BASELINE SYSTEMS

$$\mathbf{e}^* = \arg\max p(\mathbf{e}|\mathbf{f}) = \arg\max_e \{exp(\sum_i \lambda_i h_i(\mathbf{e},\mathbf{f}))\}$$

- Bilingual Ngram Translation Model [Marino et al, CL'06] (*TALPtuples*)
  - 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 funcions: 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 INTERPOLATION (POST-EVALUATION)

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

3.2 PUNCTUATION RESTORATION (PRIMARY)

- We embeded punctuation restoration in the main translation step.
  SRC: $w_1\ w_2\ w_3\ .\ \rightarrow$ <PUNC> $w_1\ w_2\ w_3$ <PUNC>
  TRG: $w_1\ w_2\ w_3\ .\ \rightarrow\ .\ w_1\ w_2\ w_3\ .$

3.3 EXPERIMENTS

- MADA+TOKAN system for disambiguation and tokenization.
- The out-of-domain was a 130K-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: *TALPtuples* system, configured to use the bilingual TM of order 4, 4-gram target-side LM and 4-gram POS target-side LM. It includes SMR with 100 statistical classes.
- Post-evaluation system: the *TALPphrases* MOSES-based system enhanced with the punctuation marks repetition and interpolation technique.

Track	System	BLEU	METEOR	Average
CRR	Union (Post-evaluation)	0.5223	0.6809	0.6016
CRR	Supplied 1 (Primary submission)	0.5263	0.6848	0.6055
CRR	Interpolation (Post-evaluation)	0.5446	0.6974	0.6210
CRR	Supplied 2 (Secondary submission)	0.4976	0.6807	0.5892
ASR	Union (Post-evaluation)	0.4379	0.6262	0.5320
ASR	Supplied 1 (Primary submission)	0.4352	0.6288	0.5320
ASR	Interpolation (Post-evaluation)	0.4562	0.6385	0.5473
ASR	Supplied 2 (Secondary submission)	0.4300	0.6292	0.5296

4 CHINESE-(ENGLISH)-SPANISH PIVOT TRANSLATION

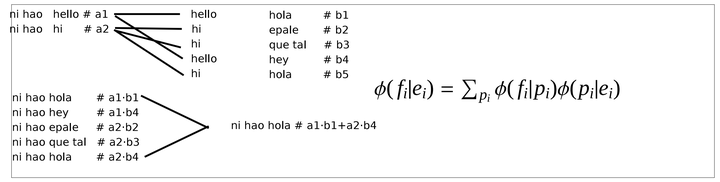
4.1 SYSTEM CASCADE (PRIMARY)

- 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, 2004]

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:



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.

Track	System	BLEU	METEOR	Average
CRR	Primary	0.3878	0.3358	0.3618
CRR	Secondary	0.3455	0.3084	0.3270
ASR	Primary	0.3513	0.3068	0.3291
ASR	Secondary	0.3063	0.2828	0.2946

5 CHINESE-TO-SPANISH DIRECT TRANSLATION

5.1 EXPERIMENTS

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

Track	System	BLEU	METEOR	Average
CRR	Primary	0.2677	0.2901	0.2789
CRR	Secondary	0.2911	0.3007	0.2959
ASR	Primary	0.2433	0.2715	0.2574
ASR	Secondary	0.2684	0.2792	0.2783

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 (phrase probabilities combination), however there is still room for improvement on phrase table pruning.
- Chinese-Spanish: Although the direct Chinese-Spanish phrase-based system performed better than the TALPtuple 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.