The TALP&I2R SMT Systems for IWSLT 2008

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

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

\texttt{\{khalilov|mruiz|carloshq|adrian|adolfohh|canton|rbanchs\}@talp.upc.edu \{bxchen|mzhang|aaiti|hli\}@i2r.a-star.edu.sg}

\textbf{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.

\section{BASELINE SYSTEMS}

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.

\subsection{PUNCTUATION RESTORATION (PRIMARY)}

We embedded punctuation restoration in the main translation step.

\subsection{EXPERIMENTS}

We performed a 200-best of possible Spanish translations for each Chinese-Spanish phrase. The final phrase probabilities were calculated as follows:

\begin{align*}
\phi (f|e) &= \sum_w \phi(f|w) \phi(w|e)
\end{align*}

\section{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.

\section{CHINESE-TO-SPANISH DIRECT TRANSLATION}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Track & System & BLEU & METEOR \\
\hline
\hline
Primary & 0.3878 & 0.6368 & 0.7047 & 0.3048 & 0.2961 \\
Secondary & 0.3423 & 0.5736 & 0.6223 & 0.2711 & 0.2626 \\
\hline
\end{tabular}
\caption{Chinese-to-Spanish direct translation results.}
\end{table}

\section{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 TALPuple 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.