The TALP&I2R SMT Systems for IWSLT 2008

Maxim Khalilov[1], Marta R. Costa-jussà[1], Carlos A. Henríquez Q.[1], José A.R. Fonollosa[1],
Adolfo Hernández H.[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

This paper gives a description of the statistical machine translation (SMT) systems developed at the TALP Research Center of the UPC (Universitat Politècnica de Catalunya) for our participation in the IWSLT'08 evaluation campaign. We present Ngram-based (**TALPtuples**) and phrase-based (**TALPphrases**) SMT systems. The paper explains the 2008 systems' architecture and outlines translation schemes we have used, mainly focusing on the new techniques that are challenged to improve speech-to-speech translation quality. The novelties we have introduced are: improved reordering method, linear combination of translation and reordering models and new technique dealing with punctuation marks insertion for a phrase-based SMT system.

This year we focus on the Arabic-English, Chinese-Spanish and pivot Chinese-(English)-Spanish translation tasks.

1. Introduction

TALP-UPC Ngram-based Machine Translation (MT) has proved to be a competitive alternative to state-of-the-art systems in previous evaluation campaigns, as shown in [1, 2]. One of the most significant distinctions of the Ngram-based translations from phrase-based systems lies in the different representation of bilingual units. It leads to a strong requirement of a certain reordering strategy implemented with probabilistic distortion model able to cope with middle- and long-distance dependencies.

Our ongoing efforts are mainly dedicated to finding the best way to reorder the source side of the bilingual corpus aiming to decrease the divergences in word order of the source and target languages, and, consequently, to reduce the size of bilingual units that the Ngram-based translation systems operates with. This is especially important when the translation is performed between pairs of languages with non-monotonic word order, like Arabic and English, or Chinese and Spanish.

Another promising way to improve the quality of MT output is to involve additional out-of-domain parallel information into bilingual modelling. Inspired by the results presented in [3], we interpolate a principal translation model (TM) with a secondary one, adjusting the weight coefficients according to the corresponding monolingual language models. To the best of our knowledge, so far no attempts have been made to linearly combine the TMs. Unfortunately, we did not have time to include the results of TM interpolation technique into the evaluation submission, but we present the post-evaluation results in the paper.

Apart from the classical Arabic-English translation, this year we have participated in a new comparative task: direct Chinese-Spanish translation versus pivot Chinese-(English)-Spanish translation.

2. Ngram-based Machine Translation system

Here we briefly describe the baseline Ngram-based translation system that coincides with the MT system used in the IWSLT'07 campaign, as well as specific novel techniques implemented for the IWSLT'08 evaluation.

Our translation system implements a log-linear model in which a foreign language sentence $f_1^J = f_1, f_2, ..., f_J$ is translated into another language $e_1^I = e_1, e_2, ..., e_I$ by searching for the translation hypothesis $\hat{e}_1^I$ maximizing a log-linear combination of several feature models [4]:

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.

The Ngram-based approach regards translation as a stochastic process maximizing the joint probability $p(f, e)$, leading to a decomposition based on bilingual n-grams, so-called *tuples*, that are extracted from a word-to-word alignment (performed with GIZA++ tool[1] and generated by *grow-diag-final* method [5]).

Given a certain word-aligned parallel corpus, tuples are extracted according to the following constraints [6]:

- a monotonic segmentation of each bilingual sentence pair is produced

- no word in a tuple is aligned to words outside of it

- no smaller tuples can be extracted without violating the previous constraints

As mentioned above, dealing with pairs of languages with non-monotonic word order, a certain reordering strategy is required to extract more reusable units (less sparse). The method that we used in this evaluation is detailed below.

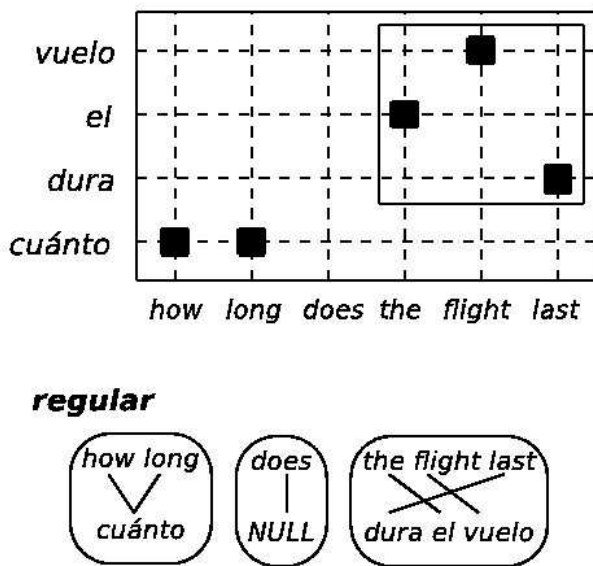Figure 1 shows an example of tuple monotonic extraction (Spanish-English).

Figure 1: *Example of tuples extraction.*

2.1. Translation model

The core part of the system following Ngram-based approach is a TM, which is based on *tuples* extracted from a word-to-word alignment. In contrast to phrase-based models, our TM is estimated as a standard n-gram model of a bilingual language expressed in *tuples*. In this way, it approximates the joint probability between source and target

---

[1] http://code.google.com/p/giza-pp/

languages capturing bilingual context, as described by the following equation:

$$p(S, T) = \prod_{k=1}^{K} p((\tilde{s}, \tilde{t})_k | (\tilde{s}, \tilde{t})_{k-N+1}, ..., (\tilde{s}, \tilde{t})_{k-1}) \quad (1)$$

where s refers to source, t to target, and $(\tilde{s}, \tilde{t})_k$ to the kth tuple of a given bilingual sentence pair segmented in K tuples.

The bilingual TM actually constitutes an n-gram-based language model (LM) of tuples, which approximates the joint probability between the languages under consideration and can be seen here as a LM, where the language is composed by tuples.

2.2. Feature functions

Apart from the TM, TALP-UPC translation system implements a log-linear combination of six additional feature models:

- a **target LM** (a model of target-side words);

- a **Part-of-Speech (POS) target LM** (a model of target-side tags);

- a **word bonus model** (is used to compensate the system's preference for short output sentences);

- a **source-to-target lexicon model** and a **target-to-source lexicon model** (the models using word-to-word IBM Model 1 probabilities to estimate the lexical weights for each tuple in the translation table);

- a **POS source LM** (a model of source-side tags, supporting reordering process);

2.3. MARIE decoder

As decoder, we use MARIE [7], a beam-search decoder developed at TALP Research Center which taking the previous models into account. For efficient pruning of the search space, *threshold pruning*, *histogram pruning* and *hypothesis recombination* are used.

MARIE admits a weighted reordering graph (distortion of source-side words order), generated by the statistical machine reordering algorithm as described in Section 2.5.

2.4. Feature weights optimization

Given the development set and references, the log-linear combination of weights was adjusted using a simplex optimization method (with the optimization criteria of the highest BLEU score) and an n-best re-ranking just as described in *http://www.statmt.org/jhuws/*. This strategy allows for a faster and more efficient adjustment of model weights by means of a double-loop optimization, which provides reduction of the number of translations that should be carried out.

2.5. Statistical Machine Reordering

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. 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.

The details are described in [8] and [9].

2.6. Translation models interpolation

During the post-evaluation period we have implemented a TM interpolation strategy following the ideas proposed in [3], where the authors present a promising technique of target LMs linear interpolation. These findings open the way to involve additional monolingual information into the translation process, and also gives a motivation to interpolate the translation and reordering tables in a linear way.

Due to a small amount of available in-domain data (IWSLT training material), we have used an out-of-domain 130K-line subset from the Arabic News, English Translation of Arabic Treebank and Ummah LDC parallel corpora (*VIOLIN*) [10] to increase the final translation and reordering tables. Both corpus statistics can be found in table 1.

Instead of time-consuming iterative TM reconstruction and using the highest BLEU score as an maximization criteria, we adjust the weights as a function of the lowest perplexity estimated by the corresponding interpolated combination of the target-side LMs and generalize the optimization results on the interpolated translation and reordering models.

The word-to-word alignment was obtained from the joint database (IWSLT + VIOLIN). Then, we separately computed the translation and reordering tables corresponding to the IWSLT and VIOLIN parts of the joint alignment. The final tables, as well as the final target LM were obtained using linear interpolation. The weight coefficients (IWSLT weight = 0.95, VIOLIN weight = 0.05) were selected using a minimum perplexity criterion estimated on the corresponding interpolated combination of the target-side LMs.

3. Phrase-based Machine Translation

In this section we present a phrase-based MT system that was used in the evaluation. This system is based on the well-known MOSES[2] toolkit, which is nowadays considered as a state-of-the-art SMT system [11]. The training and weights tuning procedures are explained in details in the above-mentioned publication, as well as, on the MOSES web page: *http://www.statmt.org/moses/*.

3.1. Punctuation restoration

We decided to embed punctuation restoration in the main translation step. For this purpose we preprocessed the training corpus as follows:

1. Source sentences: we added a `<PUNC>` tag at the beginning of each sentence, we replaced final punctuations marks (points and questions marks) with another `<PUNC>` tag, and we removed any other punctuation marks.

2. Target sentences: we repeated the final punctuation mark at the begin of each sentence.

The resulting preprocessed training corpus is used to train a standard SMT system (w_i stands for the i-th word).

SRC: $w_1\ w_2\ w_3$. $\rightarrow$ `<PUNC>` $w_1\ w_2\ w_3$ `<PUNC>`
TRG: $w_1\ w_2\ w_3$. $\rightarrow$. $w_1\ w_2\ w_3$.

During the actual translation of unpunctuated test sentences we add the `<PUNC>` tag at the beginning and at the end of each sentence. The trained TM along with the target LM and the other features serves to restore the corresponding final/initial punctuation mark translating each `<PUNC>` tag.

The rest of punctuation marks can also be restored as any other words included in the target side of translation units.

Note that the preprocessing of the target data follows the IWSLT 2008 suggestions[3], but no additional target LM is needed in this case. After translation, the same suggested postprocessing scheme is applied: the last punctuation mark is replaced with the first one and the first punctuation mark is then removed.

4. Experiments

4.1. Arabic to English translation

The first run we have participated was a Basic Traveling Expression Corpus (BTEC) Arabic to English translation task. The model weights were tuned with the 2006 development corpus (Dev6), containing 489 sentences and 6 reference translations and the 2002 development set (500 sentences and 16 reference translations) was used as an internal test, according to which we take a decision about better or worse system performance.

4.1.1. Arabic data preprocessing

We used a similar approach to that shown in [12], namely the MADA+TOKAN system for disambiguation and tokenization. For disambiguation only diacritic unigram statistics were employed. For tokenization we used the D3 scheme with -TAGBIES option. The scheme splits the following set of enclitics: w+, f+, b+, k+, l+, Al+ and pronominal enclitics. The -TAGBIES option produces Bies POS tags on all taggable tokens.

	IWSLT		VIOLIN	
	Arabic	English	Arabic	English
Sentences	24.45 K	24.45 K	130.59 K	130.59 K
Words	170.24 K	188.54 K	4.12 M	4.44 M
Average sentence length	6.96	7.71	31.52	34.01
Vocabulary	10.89 K	6.92 K	72.9 K	65.9 K

Table 1: *The main and additional basic corpora statistics.*

4.1.2. Primary submission

As a primary system we submitted the *TALPphrases* MOSES-based system enhanced with the punctuation marks repetition technique. True case restoration, required to be done on the postprocess step, was performed with the MOSES package, using a standard *recase.perl* script.

4.1.3. Secondary submission

Our secondary submission was the *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 as described in Section 2.5 with 100 statistical classes.

For this system configuration we used a strategy for restoring punctuation and case information as proposed on the IWSLT'08 web page, using standard SRI LM[13] tools: *disambig* to restore case information and *hidden-ngram* to insert missing punctuation marks.

4.1.4. Post-evaluation experiments and official evaluation results

After the systems submission we performed experiments interpolating translation and reordering tables using the weights that cause the minimal perplexity value for the interpolated target-side LM, as described in Section 2.6. The final tables were passed to the primary (MOSES-based) system. For comparison, we also estimated a standard TM from the union of the IWSLT and VIOLIN corpora.

The official submission and post-evaluation results for the ASR and CRR Arabic-English translation tasks can be found in table 2. Evaluation conditions were case-sensitive and with punctuation marks considered.

Consecutive union of the in-domain and out-of-domain corpora (*"Union"*) leads to slightly worse results for the CRR track and shows almost the same performance as the system which use the IWSLT parallel corpus solely in the MOSES-based system (*"Supplied 1"*) for the ASR track.

The system based on the weighted and merged TM (*"Interpolation"*) outperforms BTEC-only system by 1.8 BLEU points and 1.2 METEOR points for the CRR track and by 2.1 BLEU points and by about 1 METEOR points for the ASR track measured on the official evaluation test set.

"Supplied 2" line stands for the results obtained with the *TALPtuples* system as described in sub-section 4.1.3.

4.2. Chinese-(English)-Spanish pivot translation

Our participation in this task is the result of a joint contribution between I2R (Institute for Infocomm Research) and UPC. We followed two different strategies for the primary and secondary runs. In both cases the I2R team built a Chinese-to-English SMT system and the UPC team was responsible for an English-to-Spanish SMT system.

Both Machine Translation Systems were based on MOSES open source package [11]. IBM word reordering constraints [14] were applied during decoding to reduce the computational complexity. The other models and feature functions employed by MOSES decoder were:

- TM(s), direct and inverse phrase/word based TM.

- Distortion model, which assigns a cost linear to the reordering distance, while the cost is based on the number of source words which are skipped when translating a new source phrase.

- Lexicalized word reordering model [15].

- Word and phrase penalties, which count the number of words and phrases in the target string.

- Target-side LM.

The TM and reordering model were trained using the standard MOSES tools. Weights of feature functions were tuned by using the optimization tools from the MOSES package. The search operation was accomplished by MOSES decoder.

The experiments with the Chinese to English MT were carried out on the BTEC Chinese-English data [16] augmented with HIT-corpus[4], Olympic-corpus[5] and PKU-corpus[6] from Chinese LDC.

20K BTEC sentence pairs were supplied for the IWSLT 2008 evaluation campaign. HIT corpus contains 132K sentence pairs in total, and is known as a multi-source Chinese-English parallel corpus; Olympic corpus has 54K bilingual sentences mainly from sport and travelling domains; while PKU-corpus has about 200K parallel phrases and is considered as a domain-balanced corpus. Besides, the English part of the Tanaka corpus[7] was used as a complementary training

---

[4] http://mitlab.hit.edu.cn/index.php/resources
[5] http://www.chineseldc.org/EN/index.htm 2004-863-008
[6] http://www.chineseldc.org/EN/index.htm CLDC-LAC-2003-006
[7] http://www.csse.monash.edu.au/~jwb/tanakacorpus.html

Track	System	BLEU	METEOR	(BLEU+METEOR)/2	NIST
CRR	Union (Post-evaluation)	0.5223	0.6809	0.6016	8.5253
CRR	Supplied 1 (Primary submission)	0.5263	0.6848	0.6055	8.5940
CRR	Interpolation (Post-evaluation)	0.5446	0.6974	0.6210	8.8772
CRR	Supplied 2 (Secondary submission)	0.4976	0.6807	0.5892	8.7421
ASR	Union (Post-evaluation)	0.4379	0.6262	0.5320	7.2878
ASR	Supplied 1 (Primary submission)	0.4352	0.6288	0.5320	7.2808
ASR	Interpolation (Post-evaluation)	0.4562	0.6385	0.5473	7.6113
ASR	Supplied 2 (Secondary submission)	0.4300	0.6292	0.5296	7.5862

Table 2: *Official and post-evaluation results for Arabic-English translation.*

	IWSLT'08			All additional data	
	Chinese	English	Spanish	Chinese	English
Sentences	19,972	19,972	19,972	379,065	379,065
Words	164K	182K	147K	4,834K	5,036K
Vocabulary	8,506	8,301	16,953	57,055	75,156

Table 3: *Corpus used during the Chinese-English training*

material for the target-side LM.

The I2R research group performed word segmentation for the Chinese part using ICTCLAS tools[8] developed in the ICT [17].

Table 3 reports the basic statistics of the principal and additional corpora that were used to build the Chinese-to-English SMT system. Regarding English-to-Spanish translation, no extra corpora were used.

4.2.1. Chinese-English independent results

The union of the BTEC corpus and the additional bilingual corpora allowed gaining 0.23 BLEU points for the internal test set. This impact can be seen in table 4. The Chinese-English SMT system returns the sentences with truecase and tokenized punctuation, ready to be input to the English-Spanish SMT.

	IWSLT'08	Additional data
BLEU	0.3628	0.5916
NIST	7.2417	9.4015
METEOR	0.5913	0.7148

Table 4: *Results for Chinese-English translation.*

4.2.2. English-Spanish independent results

As mentioned before, no additional corpus was used for the English-to-Spanish system. The input was considered to be in true case, tokenized and with punctuation marks. Contractions like "we'll" and "you're" were split as "we 'll" and

"you 're", and negations like "don't", "wouldn't" or "can't" were split as "do n't", "would n't" and "ca n't".

The output of this system was performed in accordance with the official evaluation specification, without any post-processing needed. Table 5 shows the results of the English-Spanish system trained with the BTEC corpus.

	IWSLT'08
BLEU	0.5586
NIST	9.2855
METEOR	0.6994

Table 5: *Results for English-Spanish translation.*

4.2.3. Primary submission

Our primary approach to the pivot task was *a system cascade*. 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. From that 200-best list, which is allowed for repetitions, the single-best translation was computed using a Minimum Bayes Risk (MBR) strategy as described in [18]. We used the MOSES implementation of the MBR algorithm. This strategy of 200-best list rescoring performed better than a single-best list selection for both systems, gaining 2.5 BLEU points in the development set.

---

[8]http://www.nlp.org.cn/project/project.php?proj_id=6

4.2.4. Secondary submission

As an alternative approach to the system cascade, we followed a different strategy for the secondary submission combining the phrase translation probabilities of the two language pairs (Chinese-English and English-Spanish translations) with the strategy proposed in [19] to obtain the translation probabilities for each Chinese-Spanish phrase. The final phrase probabilities are calculated as followed:

$$\phi(f_i|e_i) = \sum_{p_i} \phi(f_i|p_i)\phi(p_i|e_i) \tag{2}$$

where $\phi(f_i|e_i)$ corresponds to the translation probability of the Chinese phrase f_i given the Spanish phrase e_i, $\phi(f_i|p_i)$ stands for the translation probability of the Chinese phrase f_i given the English phrase p_i and $\phi(p_i|e_i)$ stands for the translation probability of the English phrase p_i given the Spanish phrase e_i.

It is important to mention that the English and Spanish phrases are lowercased in this system and the case information restoration process is performed on the postprocess step, following the strategy proposed for the IWSLT'08 evaluation. These two scores are supported by a Spanish LM, a word and phrase penalty feature and a distortion model which would complete the final Chinese-(English)-Spanish system. Inspired by the idea proposed in [19] which was to extend a small amount of available parallel training material for a given language pair, we tried to use these findings to skip the English LM used during the Chinese-English translation as was applied in the cascade system. The reason for skipping the LM is that an English reordering should not be needed to get a Spanish translation from a Chinese input.

4.2.5. Chinese-(English)-Spanish pivot results

Table 6 shows the official results obtained with both strategies. As can be seen, the cascade system outperforms the secondary system, despite that the secondary submission did not use an extra LM on the pivot step of the translation process. On the other hand, the lexicalized word reordering is lost with the artificial phrase pairs and we had problems with the final table size of our TM. It happened that the computation showed in equation 2 gave us a lot of wrong phrase-pairs and most of the phrases got very low probabilities (due to the multiplication factor).

We hope that an improved pruning algorithm applied to the resulting phrase table could help to achieve a robust TM which finally would perform better than the cascade approach.

4.3. Chinese to Spanish translation

For direct Chinese-Spanish task we planned to build a Ngram-based SMT system (TALPTuple), using the SMR algorithm described in section 2.5 and a phrased-based SMT system (MOSES-based), as described in section 4.2.

For this task we only used the BTEC'08 corpus, which contains about $20,000$ sentences for training and 506 sentences with 16 Spanish references for tuning the system. The basic statistics of this corpus can be seen in table 7.

4.3.1. Data preprocessing

The Chinese corpus was not preprocessed before translation: the corpus was tokenized by words and the punctuation marks were separated.

Note that the TM, as well as the LM and reordering model, was trained with punctuation marks and the official test set that did not contain this information, therefore it was preprocessed with the *hidden-ngram* tool to restore it.

The Spanish part of the corpus was lowercased and tokenized using the Freeling toolkit[20], an open source tool for language analysis. It splitted the enclitics from the Spanish verbs (*dámelo → da +me +lo*) and also generated the POS tags that were lately used to estimate a target-side POS LM and in postprocessing.

4.3.2. Data postprocessing

Once the decoding process had finished, the output of the system was still lowercased and splitted with the enclitics and the POS tags were generated.

Afterwards, a postprocess including two steps was performed: firstly, the original morphological verbs form was restored using the enclitics and POS tags information; on the next step, the case information is restored, using the *disambig* tool from SRILM following the instruction from the IWSLT'08 web-page. This postprocess was not run during the tuning step, where all the Spanish references were also tokenized, splitted with enclitics and lowercased.

4.3.3. Chinese to Spanish translation results

The official evaluation results for both systems can be seen in table 8.

As mentioned before, the TALPTuple system was nominated as the primary submission and the MOSES-based system as the secondary one.

5. Conclusions

In this year evaluation we participated in three translation tasks, collaborating with the I2R in pivot Chinese-(English)-Spanish translation tasks. This paper outlines the architecture of the submitted translation systems and summarizes the preliminary official results.

The main conclusion that can be made from our participation in the Arabic-English shared task is that the Ngram-based system is comparable with the state-of-the-art phrase-based SMT in terms of automatically evaluated accuracy for both the ASR and CRR tasks: in case of CRR track the MOSES system outperforms the tuples-based one by 3 BLEU points, but loosing in NIST score, while for the ASR

Submission	BLEU	METEOR	(BLEU+METEOR)/2	NIST
Primary ASR	0.3513	0.3068	0.3291	5.1516
Primary CRR	0.3878	0.3358	0.3618	5.7953
Secondary ASR	0.3063	0.2828	0.2946	4.0314
Secondary CRR	0.3455	0.3084	0.3270	4.6075

Table 6: *Official results for the Chinese-(English)-Spanish translation.*

	Chinese	Spanish
Sentences	19,972	19,972
Words	171K	147K
Average sentence length	8.59	7.39
Vocabulary	8,428	16,953

Table 7: *Corpus used for the Chinese-Spanish training*

Submission	BLEU	METEOR	(BLEU+METEOR)/2	NIST
Primary ASR	0.2433	0.2715	0.2574	5.2547
Primary CRR	0.2677	0.2901	0.2789	5.6833
Secondary ASR	0.2684	0.2792	0.2783	4.9303
Secondary CRR	0.2911	0.3007	0.2959	5.3240

Table 8: *Official results for Chinese-Spanish translation.*

run the difference in BLEU and METEOR results is negligible and the Ngram-based translation is evaluated slightly higher in terms of NIST metric.

For the Chinese-(English)-Spanish pivot task the system cascade architecture demonstrates better results than the alternative (phrase probabilities combination), however there is still room for improvement on phrase table pruning. 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.

Future work is to be conducted to apply the promising TM interpolation strategy to the Ngram-based SMT.

6. Acknowledgments

7. References

[1] M. Paul, "Overview of the IWSLT 2006 Evaluation Campaign," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT'06)*, Kyoto, Japan, 2006, pp. 2–15.

[2] C. S. Fordyce, "Overview of the IWSLT 2007 evaluation campaign," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT'07)*, Trento, Italy, 2007, pp. 1–12.

[3] H. Schwenk and Y. Estève, "Data selection and smoothing in an open-source system for the 2008 NIST machine translation evaluation," in *Proceedings of the Interspeech'08*, Brisbane, Australia, 2008, p. to appear.

[4] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. R. Fonollosa, M. R. Costa-jussà, and M. Khalilov, "UPC's bilingual n-gram translation system," in *Proc. of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 43–48.

[5] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," in *Computational Linguistics*, vol. 29(1), 2003.

[6] J. M. Crego, J. B. Mariño, and A. de Gispert, "Finite-state-based and phrase-based statistical machine translation," in *Proc. of the 8th Int. Conf. on Spoken Language Processing (ICSLP'04)*, Jeju, Korea, October 2004, pp. 37–40.

[7] J. M. Crego, J. B. Mariño, and A. de Gispert, "A ngram-based statistical machine translation decoder," in *Proc. of the Interspeech'05*, Lissbon, Portugal, September 2005, pp. 3193–96.

[8] M. R. Costa-jussà and J. R. Fonollosa, "Statistical machine reordering," in *Empirical Methods in Natural Language Processing (EMNLP'06)*, Sydney, Australia, July 2006, pp. 70–76.

[9] M. R. Costa-jussà and J. R. Fonollosa, "Analysis of statistical and morphological classes to generate weigthed reordering hypotheses on a statistical machine translation system," in *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, June 2007, pp. 171–176.

[10] N. Habash, "Syntactic preprocessing for statistical machine translation," in *Proceeding of the Machine Translation Summit (MT-Summit)*, Copenhagen, Denmark, September 2007.

[11] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, June 2007, pp. 177–180.

[12] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proceedings of the Human Language Technology and North American Assosiation for Computational Linguistics Conference (HLT/NAACL'06)*, New York, USA, June 2006.

[13] A. Stolcke, "SRILM: an extensible language modeling toolkit." in *Proc. of the Int. Conf. on Spoken Language Processing*, Denver, CO, September 2002, pp. 901–904.

[14] A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, A. S. Kehler, and R. L. Mercer, "Language translation apparatus and method using context-based translation models," Patent 5 510 981, 1996.

[15] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh system description for the 2005 IWSLT speech translation evaluation," in *Proceedings of the Int. Workshop on Spoken Language Translation (IWSLT'05)*, Pittsburg, USA, October 2005.

[16] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in *Proceeding of LREC-2002: Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, May 2002, pp. 147–152.

[17] H. Zhang, H. Yu, D. Xiong, and Q. Liu, "HHMM-based chinese lexical analyzer ICTCLAS," in *Proc. of the 2nd SIGHAN Workshop on Chinese language processing*, Sapporo, Japan, July 2003, pp. 184–187.

[18] S. Kumar and W. Byrne, "Minimum bayes-risk decoding for statistical machine translation," in *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL'04)*, Boston, USA, May 2004, pp. 169–176.

[19] H. Wu and H. Wang, "Pivot language approach for phrase-based statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, Prague, Czech Republic, June 2007, pp. 856–863.

[20] X. Carreras, I. Chao, L. Padrò, and M. Padrò, "Freeling: An open-source suite of language analyzers," in *Proceeding of LREC-2004: Fourth International Conference on Language Resources and Evaluation*, Lissbon,Portugal, May 2004.