ITC-irst at the 2006 TC-STAR SLT Evaluation Campaign

Nicola Bertoldi, Mauro Cettolo, Roldano Cattoni, Boxing Chen, Marcello Federico

ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
via Sommarive 18, 38050 Povo di Trento, Italy
{surname}@itc.it
http://hermes.itc.it

Abstract

This paper reports on the participation of ITC-irst in the 2006 Spoken Language Translation Evaluation Campaign organized by the TC-STAR project. ITC-irst submitted runs for all translation directions, namely Spanish-to-English, English-to-Spanish and Chinese-to-English, and types of input, that is final text edition, human verbatim transcriptions and speech recognition output. Official results show that translations produced by our systems rank among the best ones. With respect to the translation systems we developed for the 2005 evaluation, BLEU scores were improved in every condition, from 17% up to 40% relative.

1. Introduction

This paper reports on the systems developed at ITC-irst for the 2006 Spoken Language Translation (SLT) Evaluation Campaign organized by the TC-STAR project[1]. ITC-irst submitted runs for all translation directions and types of input. Official results show that our systems rank among the best ones participating in the evaluation.

The paper describes the ITC-irst systems developed for all translation directions, namely Spanish-to-English, English-to-Spanish and Chinese-to-English, and input types, namely final text edition (FTE), verbatim human transcription (VHT) and speech recognition output (ASR). Performances of the 2006 systems are compared with those achieved by the systems developed for the 2005 Evaluation Campaign; comparison is provided in terms of WER, BLEU (Papineni et al., 2001) and NIST (Doddington, 2002) scores.

The paper is organized as follows. Section 2 presents the general log-linear framework of Statistical Machine Translation (SMT), and gives an overview of the phrase-based SMT architecture and all other aspects which are shared by all developed systems. Sections 3 and 4 give details peculiar to each system, such as used training data and system setting, and compare its performance against the corresponding 2005 system. Section 5 reports on the qualitative improvement observed over the outputs of the last year, showing significant examples. Finally, the paper ends with some concluding remarks.

2. Phrase-based Translation System

Given a string $\mathbf{f}$ in the source language, the goal of statistical machine translation is to search for the string $\mathbf{e}$ in the target language which maximizes the posterior distribution $\Pr(\mathbf{e} \mid \mathbf{f})$. In phrase-based translation, words are no longer the only units of translation, but they are complemented by strings of consecutive words, the phrases. By assuming a log-linear model (Berger et al., 1996; Och and Ney, 2002) and by introducing the concept of word alignment (Brown et al., 1993), the optimal translation can be searched for with the criterion:

$$\tilde{\mathbf{e}}^* = \arg\max_{\tilde{\mathbf{e}}} \max_{\mathbf{a}} \sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a}),$$

where $\tilde{\mathbf{e}}$ represents a string of phrases in the target language, $\mathbf{a}$ an alignment from the words in $\mathbf{f}$ to the phrases in $\tilde{\mathbf{e}}$, and $h_r(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a})$ $r = 1, \ldots, R$ are *feature functions*, designed to model different aspects of the translation process. The assumed translation process extends step by step the target string by covering new source positions until all of them are covered. For each added target phrase, a source phrase within $\mathbf{f}$ is chosen, and the corresponding score is computed on the basis of its position and phrase-to-phrase translation probabilities. The fluency of the added target phrase with respect to its left context is evaluated by a n-gram language model. Some exceptions are also managed: target words might be added which do not translate any source word, and some of the source words can be left untranslated.

2.1. Model Training

The resulting log-linear model embeds feature functions whose parameters are either estimated from data or empirically fixed. The scaling factors λ of the log-linear model can be estimated on a development set, by applying a *minimum error training* procedure (Och, 2003; Cettolo and Federico, 2004).

The language model feature function is estimated on unsegmented and lowercased monolingual texts.

The phrase-to-phrase probability feature is estimated from phrase-pair statistics extracted from word-aligned and lowercased parallel texts. Direct and inverse alignments are computed with the GIZA++ software tool (Och and Ney, 2000) which implements statistical models developed by (Brown et al., 1993; Och and Ney, 2000). Phrase pairs are extracted from the aligned texts by means of the algorithm described in (Chen et al., 2005). We set the maximum phrase length to 8 words.

2.2. Decoding Strategy

Figure 1 illustrates how the translation of an input string is performed by our SMT system (Chen et al., 2005). In the first stage, a beam search algorithm (decoder) computes a word graph of translation hypotheses. Hence, either the

[1] www.tc-star.org

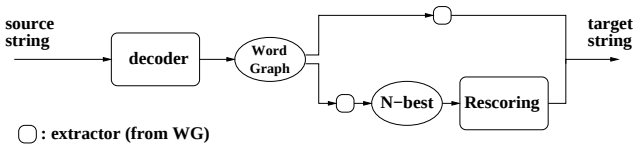

Figure 1: Architecture of the ITC-irst SMT system.

best translation hypothesis is directly extracted from the word graph and output, or an N-best list of translations is computed by means of an exact algorithm (Tran et al., 1996). The N-best translations are rescored by applying additional features and re-ranked.

The decoder exploits dynamic programming, i.e. the optimal solution is computed by expanding and recombining previously computed partial theories. Theory expansion basically follows the translation process explained above.

To cope with the large number of generated theories, a beam is used to prune out partial theories that are less promising and constraints are set to possible word reordering.

Pruning is applied on all theories covering the same set of source positions, and on all theories with the same output length.

Word re-ordering constraints are applied during translation each time a new source position is covered, by limiting the maximum number of vacant positions on the left (MVN) and the maximum distance from the left most vacant position (MVD). The two parameters MVN and MVD were set to different values according to the task, as reported later.

2.3. Feature Functions

The list of features used for translation follows. Those employed in both translation stages are marked with (1,2), while features used only by the rescoring module are marked with (2).

- Target 4-gram LM (1,2): standard language model smoothed by applying the *modified Kneser-Ney* method (Goodman and Chen, 1998).
- Fertility model (1,2): for each target phrase, it guesses the number of words of the corresponding source phrase. It is estimated from parallel texts.
- Direct phrase-based lexicon model (1,2): it consists of translation probability distributions of source phrases for each target phrase.
- Inverse phrase-based lexicon model (1,2): it provides inverted distributions with respect to the previous model.
- Direct word-based lexicon model (1,2): phrase-based translation probabilities are computed according to IBM Model 1
- Positive distortion model (1,2): this is a negative exponential model which assigns probabilities to positive position skips
- Negative distortion model (1,2): like the previous mode but for negative position skips.
- Rules extracted from dev sets (1,2): explicit phrase-to-phrase translation hints (Cettolo et al., 2005) generated on-the-fly by means of manually designed rules, given in form of regular expressions.
- Length penalty (non-normalized) (2): a simple counter of target words which should favor longer hypotheses.
- IBM Model 1 and 3 lexicons (2): these lexicons are used to rescore translation alternatives according to the IBM Model 1 formula (Brown et al., 1993). They should capture lexical co-occurrences in the source and target strings.
- Additional target n-gram LMs (2).
- n-grams in N-best (2): counts of n-grams of length 1 to 4 occurring in the N-best translations; it favors hypotheses containing popular n-grams (Chen et al., 2005).

2.4. Post-processing

Post-processing basically involves case restoration of the target language (English or Spanish), i.e. recovering word case information of proper names, words after strong punctuation, etc. For this purpose, we used the `disambig` tool[2] fed with a n-gram cased language model.

3. Spanish-to-English and English-to-Spanish Tasks

3.1. Training Data

Spanish-to-English and English-to-Spanish systems were trained on the whole European Parliament Plenary Sessions corpus (EPPS) consisting of 34M Spanish and 33M English running words.

The EPPS corpus contains so-called final text editions of the parliamentary speeches, which are more formal and syntactically correct with respect to the corresponding human transcriptions. In order to adapt the systems to the VHT and ASR conditions, we exploited monolingual training data provided for the TC-STAR ASR Evaluation Campaign. In particular, 520K and 792K running words were used for Spanish and English, respectively.

Adaptation to the VHT and ASR conditions was performed by estimating condition dependent target language models and using them as additional feature in the second stage.

3.2. System Settings

A uniform linear combination of decoder features (all weights set to 1) was used in the first step.

Strong constraints on word reordering was applied by setting MVD=MVN=1 and MVD=MVN=2 for Spanish-to-English and English-to-Spanish, respectively.

Translation model was pruned by removing all singleton phrase pairs. Moreover, for each source phrase only the most probable translations whose total probability is above a given threshold were kept in the model; the threshold was set to 0.8 and 0.9 for Spanish-to-English and English-to-Spanish, respectively. In any case, no more than 30 translations were included. After pruning, translation models contain 3.9M and 4.2M phrase pairs for Spanish-to-English and English-to-Spanish, respectively.

Concerning the second stage of the translation process, we rescored the 1000-best translation hypotheses with the additional features described in the previous section. Actually, we replaced the 4-gram LM used for the first step with

[2] `www.speech.sri.com/projects/srilm`

3 800 millones de euros	$\Longrightarrow$	3800 million euros
mil millones de euros	$\Longrightarrow$	billion euros
garriga polledo	$\Longrightarrow$	garriga polledo
verts	$\Longrightarrow$	verts

Table 1: Examples of Spanish-to-English translation rules suggested to the decoder.

	BLEU	ΔBLEU
baseline	57.91	–
+ibm1	58.17	+0.26
+ibm3	58.19	+0.28
+nbest-4gr	58.56	+0.65
+len	58.38	+0.47
+5gr-lm	58.10	+0.19

Table 2: Impact of single additional features used in the rescoring step. Figures refer to the Spanish-to-English 2005 FTE dev set.

a 5-gram LM trained on the same data. We also added a 5-gram LM trained on the development set provided for the 2006 SLT TC-STAR Evaluation. Finally, only for the VHT and ASR conditions, we used a 5-gram LM trained on the training data provided for the 2006 TC-STAR ASR Evaluation.

Feature weights were tuned empirically on the development set of the First TC-STAR SLT Evaluation by taking into account a linear combination of four translation quality measures: BLEU+4*NIST+(100-WER)+(100-PER). Automatic estimation of weights did not give any further benefit.

A small set of rules was created to suggest to the decoder good translations of numerical and currency expressions. Further rules were provided for detecting and translating the names of the members and the parties of the European and Spanish Parliaments. It is worth noticing that a list of such names is indeed a monolingual resource. Some examples of rules are given in Table 1.

Case restoration uses a 4-gram case-sensitive LM estimated on the same sample used to estimate the LM of the decoder.

3.3. Results

Performance improvements given by using two translation steps instead of one are now analyzed. Firstly, the contribution of each single additional feature is taken into account. Table 2 shows the BLEU score and the relative improvement achieved by rescoring the 1000-best list with one feature at a time. Feature weights were set to 1. Experiments refer to the 2005 development set of the Spanish-to-English FTE condition.

Table 3 compares performance of the single-step and two-step systems with respect to different translation quality measures. Comparison is provided for Spanish-to-English and English-to-Spanish translation directions on the 2005 test set. Figures show that the rescoring allows to increment Spanish-to-English translation performance for all conditions. On the contrary, no improvement was observed for the English-to-Spanish tasks: this is probably due to a less fine tuning of the rescoring module.

Finally, Table 4 reports figures which compare performance of our 2005 and 2006 systems. Comparison is done on the 2005 test set for FTE, VHT and ASR Spanish-to-English conditions. Figures show that the new system improves BLEU score over the old one by 17 to 22% relative.

CHI-ENG	parallel resources		monolingual resources
	Chinese	English	English
training	82M	88M	464M

Table 5: Statistics of the Chinese-to-English training data (running words.

4. Chinese-to-English Tasks

4.1. Preprocessing

In Chinese texts, word segmentation was performed by means of ICTCLAS, a publicly available tool developed at the Institute of Computing Technology, Beijing (Zhang et al., 2003). Then a tokenization step separates words from punctuation. A similar tokenization was applied to the target sentences; in addition, numbers written in textual form were transformed into digits and words were put in lower case. Parallel sentences were filtered out if source and target differ too much in length.

Since long parallel texts represent a problem in training word-alignment models, they were split into smaller parallel segments by means of a binary and recursive procedure. The method relies on a likelihood measure which evaluates the correspondence of a segment pair in the source and target language, respectively. Text break candidates are chosen both according to strong punctuation and segment length. For our training, the final length of parallel segments is at most 30 words.

4.2. Training Data

Table 5 provides figures on training corpora. They do not include statistics of Named Entities (LDC2003E01), NIST 2002 test set (LDC2003T17) nor UN data (LDC2004E12), which were used only in a weak manner, as explained later.

All bilingual LDC resources listed in the TCSTAR web site were used for the estimation of translation models, with the exception of the Chinese Treebank (LDC2005T01). Singleton phrase pairs were not considered for training the lexicon models with the exception of those also observed in either Named Entities or UN corpora.

As monolingual resource (English), the portion “Xinhua” and “Afe” of the “English Gigaword” corpus (LDC2003T05) were added to the allowed LDC bilingual corpora (UN corpus was not used). The decoder 4-gram LM was estimated on the data and then adapted on the development set by means of a mixture-based method (Federico and Bertoldi, 2001).

Concerning the training of additional LMs, a 5-gram LM was estimated without using Xinhua nor Afe texts, while a 3-gram LM was estimated on NIST 2002 test set.

task		1-step			2-step		
		WER	BLEU	NIST	WER (%Δ)	BLEU (%Δ)	NIST (%Δ)
Spanish-to-English	FTE	35.6	54.1	10.52	34.9 (-2.0%)	55.1 (+1.8%)	10.60 (+0.8%)
Spanish-to-English	VHT	44.4	43.7	9.35	44.0 (-0.9%)	45.3 (+3.7%)	9.44 (+1.0%)
Spanish-to-English	ASR	49.2	39.5	8.70	48.5 (-1.4%)	40.9 (+3.5%)	8.83 (+1.5%)
English-to-Spanish	FTE	39.9	49.5	9.93	40.1 (+0.5%)	49.5 (0.0%)	9.92 (-0.1%)
English-to-Spanish	VHT	48.4	40.8	9.00	48.5 (+0.2%)	41.1 (+0.7%)	9.02 (+0.2%)
English-to-Spanish	ASR	52.3	37.0	8.43	52.13 (-0.4%)	37.5 (+1.4%)	8.46 (+0.4%)
Chinese-to-English	FTE	77.5	16.0	6.02	77.7 (+0.3%)	17.5 (+9.4%)	6.10 (+1.3%)
Chinese-to-English	VHT	81.0	14.7	5.82	80.6 (-0.5%)	16.4 (+11.6%)	5.94 (+2.1%)
Chinese-to-English	ASR	81.0	14.5	5.68	80.6 (-0.5%)	16.1 (+11.0%)	5.78 (+1.7%)

Table 3: ITC-irst's 2006 systems: impact of the rescoring module on several translation tasks of the TC-STAR 2005 campaign. Scores refer to case-sensitive evaluation.

task		system 2005			system 2006		
		WER	BLEU	NIST	WER (%Δ)	BLEU (%Δ)	NIST (%Δ)
Spanish-to-English	FTE	40.9	47.0	9.54	34.9 (-14.7%)	55.1 (+17.2%)	10.60 (+11.1%)
Spanish-to-English	VHT	51.1	37.2	8.31	44.0 (-13.9%)	45.3 (+21.8%)	9.44 (+13.6%)
Spanish-to-English	ASR	54.6	33.8	7.83	48.5 (-11.2%)	40.9 (+21.0%)	8.83 (+11.3%)
Chinese-to-English	FTE	82.3	12.6	5.36	77.7 (-5.7%)	17.5 (+38.9%)	6.10 (+13.8%)
Chinese-to-English	VHT	83.6	12.0	5.37	80.6 (-3.6%)	16.4 (+36.7%)	5.94 (+10.6%)
Chinese-to-English	ASR	83.7	11.5	5.20	80.6 (-3.7%)	16.1 (+40.0%)	5.78 (+11.2%)

Table 4: Performance of ITC-irst's 2005 and 2006 systems on several translation tasks of the TC-STAR 2005 campaign. Scores refer to case-sensitive evaluation.

4.3. System Settings

Phrase-to-phrase distributions actually exploited during translation are limited by the following two thresholds: for each source phrase, only translations are kept whose total probability mass sums up to 0.99, and in any case no more than 30 target phrases are considered.

The decoder uses scaling factors λ's estimated by maximizing the BLEU score over the NIST 2003 test set. On the contrary, λ's employed for re-ranking 5000-best lists are "flat", i.e. they were not estimated but empirically fixed to 1 with few exceptions.

Both the two parameters MVN and MVD were set to 5.

Case restoration uses a 3-gram cased LM estimated on the same sample used for the decoder LM.

4.4. Results

Since this year's system employs a rescoring module which was not available in the 2005 evaluation, it is interesting to highlight its contribution to the global performance. Table 3 shows the improvement by rescoring 5000-best translations provided by the decoder with respect to the first best output by the decoder. In terms of BLEU score, the second stage yields a 10-11% relative increase.

By comparing the 2005 and 2006 decoders in terms of BLEU score the new decoder improves performance by 22%-27% relative (column "1-step" of Table 3 vs. column "system 2005" of Table 4).

Finally, Table 4 reports figures which compare performance of 2005 and 2006 systems. Comparison is done on the 2005 test set for FTE, VHT and ASR conditions. BLEU score improves from 37% to 40% relative.

5. Discussion

Table 6 lists examples of actual translations into English from FTE Spanish and Chinese inputs. Outputs were generated by the 2005 decoder, by the 2006 decoder and by the complete 2-stage 2006 system.

Results presented in the previous sections put in evidence that the decoder has been significantly improved with respect to the first evaluation and that a further important gain is assured by the rescoring module, especially for the Chinese-to-English direction.

Better quality of the decoder essentially derives from better models which enhance both lexical choice and fluency of translations:

- **Lexical Choice**: In the Spanish-to-English example n. 1, the proper name was missed by the 2005 decoder; in the second example, the verb was wrong: both errors are recovered by the 2006 decoder.
  In the Chinese-to-English example n. 2, the new decoder properly outputs "future" instead of "coming" and does not miss "seriously". In the example n. 4, it does not mistake the proper name; in the example n. 5, the week day is properly translated.
- **Fluency**: In the Spanish-to-English example n. 3, the new decoder translates properly "Programa de La Haya" , while the old one does not.
  In the Chinese-to-English example n. 1, the 2005 decoder mismatches the concordance of the verb time ("was") and the adverb ("now"), error corrected by the 2006 decoder.

The rescoring module is able to improve the translation

quality too:

- **Lexical Choice**: In the Chinese-to-English example n. 1, the decoder chose the pronoun "who", probably deceived by the previous "guard", but the rescoring replaced it with the appropriate adverb "where".
- **Fluency**: In the Spanish-to-English examples n. 4 and 5, rescoring improves the fluency of the decoder output: in the first case, it "reorders" the sequence "difference European"; in the second case, it "deletes" the article coherently with the choice of not using it before "Social cohesion".
  In the Chinese-to-English example n. 1, it is able to even recover the meaning; also in the example n. 3 the fluency is improved such that the sentence is more understandable.

Of course, it may happen that the rescoring ranks as first a translation which degrades the decoder output, like in the Chinese-to-English example n. 2.

6. Conclusions

For the second TC-STAR evaluation campaign, ITC-irst developed systems for all translation directions and types of input.

We significantly improved the systems employed in the first evaluation campaign.

First of all, we implemented the rescoring module. This requires a decoder able to output multiple translation hypotheses. In fact, given a source sentence, our decoder generates a word graph from which a list of N-best translation hypotheses is extracted. The list is then enriched by scores from additional features and re-ranked: the resulting highest scoring entry of the list is provided as result.

Thanks to the rescoring module, the quality of the system improved, especially for the Chinese-to-English system. For the other two directions, the gain was smaller (Spanish-to-English) or not observed at all (English-to-Spanish).

Results presented in this report show that most of the improvement comes from the decoder. All systems have benefited from the use of three lexicon models during decoding (the previous version employed only the direct phrase-to-phrase translation model) and from a careful tuning of beam-search parameters. Moreover, concerning the Chinese-to-English task, models were estimated on more and cleaner training data.

Finally, a significant gain derived from the replacement of the case restoration module based on maximum-entropy with one based on cased n-gram LMs.

Official results showed that our submissions ranked always among the top ones.

Acknowledgments

This work has been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

7. References

A. Berger, S.A. Della Pietra, and V.J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–313.

M. Cettolo and M. Federico. 2004. Minimum error training of log-linear translation models. In *Proc. of IWSLT*.

M. Cettolo, M. Federico, N. Bertoldi, R. Cattoni, and B. Chen. 2005. A Look inside the ITC-irst SMT System. In *Proc. of the tenth MT-Summit*, Phuket, Thailand.

B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico. 2005. The ITC-IRST SMT system for IWSLT-2005. In *Proc. of IWSLT*.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of the ARPA Workshop on HLT*, San Diego, CA.

M. Federico, and N. Bertoldi 2001. Broadcast News LM Adaptation using Contemporary Texts. In *Proc. of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark.

J. Goodman and S. Chen. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.

F.J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proc. of the 38th Meeting of the ACL*, Hong Kong, China.

F.J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for stati stical machine translation. In *ACL02: Proceedings of the 40th Meeting of the ACL*, PA, Philadelphia.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41st Meeting of the ACL*, Sapporo, Japan.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Research Report RC22176, IBM Research Division, Thomas J. Watson Research Center.

B.H. Tran, F. Seide, and V. Steinbiss. 1996. A word graph based n-best search in continuous speech recognition. In *Proc. of ICLSP*, Philadelphia, PA.

H.-P. Zhang, Q. Liu, X.Q. Cheng, H. Zhang, and H.-K. Yu. 2003. Chinese lexical analysis using hierarchical hidden Markov model. In *Proc. of the Second SIGHAN Workshop on Chinese Language Processing*.

Spanish-to-English examples		
1	ref1	Mister Piebalgs , Latvia's new candidate
	ref2	Mr. Piebalgs , Latvia's new candidate
	2005	Mr , the new candidate of Latvia
	2006-1	The Mr. Piebalgs , the new candidate of Latvia
	2006-2	Mr. Piebalgs , the new candidate of Latvia
2	ref1	Although these coarse words were roundly rejected
	ref2	Though those blunt words have been widely refused
	2005	Although these crude words have been widely accepted
	2006-1	Although these crude words have been widely rejected
	2006-2	Although these crude words have been widely rejected
3	ref1	although they went unrecognised in The Hague Programme
	ref2	though they have not been taken up in The Hague Programme
	2005	although there have been included in the programme of the Hague
	2006-1	although there have been included in the Hague Programme
	2006-2	although there have been included in the Hague Programme
4	ref1	This is what the European difference should signify .
	ref2	This should be the European difference .
	2005	This should be the difference in Europe .
	2006-1	This should be the difference European .
	2006-2	This should be the European difference .
5	ref1	Social cohesion and sustainability
	ref2	The social cohesion and the sustainability
	2005	Social cohesion and the sustainability
	2006-1	Social cohesion and the sustainability
	2006-2	Social cohesion and sustainability

Chinese-to-English examples		
1	ref1	At present , he is being kept under close watch in Rome since his arrest last month .
	ref2	At present he is under tight supervision in Rome where he was arrested last month .
	2005	He was now in Rome under tight guard , where he was arrested last month .
	2006-1	He is currently in Rome under tight guard , who last month where he was arrested .
	2006-2	He is currently in Rome under tight guard , where he was arrested last month .
2	ref1	Therefore , I believe that in future the Democratic Progressive Party should face the matter seriously .
	ref2	So I believe future DPP should treat this problem seriously .
	2005	Then I think that in the coming of the DPP have to face this problem.
	2006-1	I think that in future , the DPP must seriously address this problem .
	2006-2	I think that in the future of the Democratic Progressive Party should seriously address this problem .
3	ref1	But there will be no fundamental difference of national identification .
	ref2	But there will be no radical disagreement in recognition of a state ,
	2005	However , in recognition of the country has no fundamental differences .
	2006-1	But the national identification is concerned there are no fundamental differences .
	2006-2	However , there is no fundamental differences in national identification .
4	ref1	Guo Zhengliang says that the difference of the two parties...
	ref2	Kuo Jeng-liang said , the disagreement of the two Parties ...
	2005	Kuo Cheng-Liang , said the 2 parties to the differences ...
	2006-1	Guo Zhengliang said that the 2 parties have no differences ...
	2006-2	Guo Zhengliang said that the 2 parties have no differences ...
5	ref1	Keizo Obuchi ... in Hanoi , capital of Viet Nam , on Wednesday .
	ref2	Premier Keizo Obuchi ... on Wednesday in Hanoi , the capital of Vietnam .
	2005	Keizo Obuchi was the third Monday in Vietnam's capital Hanoi ...
	2006-1	Keizo Obuchi was Wednesday in Vietnam's capital Hanoi ...
	2006-2	Keizo Obuchi was Wednesday in Vietnamese capital of Hanoi ...

Table 6: Examples of Spanish and Chinese FTE inputs translated into English by the (i) 2005 system, (ii) 2006 decoder and (iii) 2-stage 2006 system. References are also provided.