The ITC-irst SMT System for IWSLT 2006

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

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
via Sommarive 18, 38050 Povo (Trento), Italy
{boxing,cattoni,bertoldi,cettolo,federico}@itc.it

Abstract

This paper reports on the participation of ITC-irst to the evaluation campaign of the International Workshop on Spoken Language Translation 2006. Our two-pass system is the evolution of the one we employed for the 2005 campaign: in the first pass, an N-best list of translations is generated for each source sentence by means of a beam-search decoder; in the second pass, N-best lists are rescored and reranked exploiting additional feature functions. Main updates brought to the 2005 system involve novel additional features which are here described. Results on development sets are analyzed and commented.

1. Introduction

In this paper, we report on the participation of ITC-irst to the evaluation campaign of the International Workshop on Spoken Language Translation 2006.

We submitted runs under the *Open Data* conditions for all the language pairs: Arabic-to-English, Chinese-to-English, Japanese-to-English and Italian-to-English. For each language pair, we performed translations of both manual transcriptions and first-best transcriptions provided automatically by a speech recognizer.

The statistical machine translation (SMT) system developed for the evaluation is an evolution of the one employed for the 2005 IWSLT campaign [1]. The main update is the introduction of some new additional features into the rescoring module: some of them are well known; others, like modeling word reordering phenomena, are new.

The paper is organized as follows. Section 2 briefly describes the ITC-irst phrase-based SMT system, spending some more words on new additional features. In Section 3, experimental setups of the evaluation campaign runs and results are presented and discussed. A conclusion section ends the paper.

2. System Description

ITC-irst SMT system [1] implements a log-linear model and features a two-step decoding strategy. In the first pass, a dynamic programming beam search algorithm generates N-best translation hypotheses for each source sentence. In the second pass, a larger set of local and global feature functions is employed to re-score and re-rank the N-best lists. The resulting top-scored entry of each list is finally returned as best translation hypothesis.

2.1. Rescoring Models

The feature functions and search constraints adopted for decoding are quite standard: phrase and word translation models, 4-gram language model, fertility model, IBM reordering constraints, beam search. A detailed description is provided in [1, 2]. On the other hand, SMT systems often differ a lot in the models employed for rescoring N-best candidates. Here the list of those we apply:

- direct and inverse IBM model 1 and 3 lexicons, over all possible alignments
- competitive linking algorithm (CLA) lexicon score [3], over all possible alignments. Briefly, the CLA computes an association score between all possible word pairs within the parallel corpus, and then applies a greedy algorithm to compute the best word-alignment for each sentence pair
- question feature, i.e. a binary feature which triggers when text ends with a question mark and starts with one of the typical starting words of question sentences found in training data
- frequency of its n-grams (n=1,2,3,4) within the N-best translations
- ratio between the target and source length
- 2,3,5-grams target LMs
- n-gram posterior probabilities within the N-best translations [4]
- sentence length posterior probabilities [4]
- word/block reordering probabilities (Section 2.2)
- ratio of the source length and the number of source phrases (Section 2.3)

The first six feature functions were used in our system in IWSLT-2005. The two posterior probabilities represent a

Table 1: Statistics of training, development and testing data used for the IWSLT 2006 Open Data condition.

		Chinese	Japanese	English	Arabic	Italian	English
Train Data	Sentences	39,953			19,972		
	Running words	347K	392K	363K	180K	176K	182K
	Vocabulary	11,439	12,667	9,938	15,888	10,160	7,326
Dev1 (Text)	Sentences	489		489×7	489		489×7
	Running words	5,094	5,840	39,386	5,108	4,976	39,386
Dev2 (Read Speech)	Sentences	489		489×7	489		489×7
	Running words	5,086	5,992	39,386	5,237	4,937	39,386
Test1 (Text)	Sentences	500		–	500		–
	Running words	5,506	6,407	–	5,692	5,793	–
Test2 (Read Speech)	Sentences	500		–	500		–
	Running words	5,478	6,617	–	5,829	5,514	–
Test3 (Spontaneous Speech)	Sentences	500	–	–	–	–	–
	Running words	5,416	–	–	–	–	–

Table 2: Preprocessing steps applied to languages; an "x" means that the preprocessing step is performed.

Preprocessing	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng		Ita-to-Eng	
	Chinese	English	Japanese	English	Arabic	English	Italian	English
tokenization	x	x	x	x	x	x	x	x
txt-to-digit	x	x	–	–	–	–	x	x
lower-casing	–	x	–	x	–	x	x	x

refinement of the counts of n-grams in N-best lists we employed for the 2005 campaign. The last two features are new and are presented in the following.

2.2. Word/Block Reordering

In the companion paper [5], the use of rules for modeling word reordering phenomena in phrase-based SMT is proposed. Reordering rules consist of two sides: the left-hand-side (*lhs*), which is a word-based pattern, and the right-hand-side (*rhs*), which corresponds to a possible reordering of that pattern. Different rules can share the *lhs*, because the same pattern can be reordered in more than one way. Rules are automatically extracted from word aligned training data and weighted according to observed statistics.

Rules can reorder sequences of single words or a pair of blocks of words. A block is a sequence of source words which are always aligned to consecutive positions in an aligned parallel corpus; null alignments are not considered.

Once reordering rules are available, they can be applied to each input sentence. Currently, we use them in the rescoring stage as additional feature function as sketched in the following.

Given a source sentence, a list of N-best possible translations and the corresponding alignments, the rules whose *lhs* matches totally or partially the source string are listed. Now, for each entry in the N-best list, among all the rules matching the *lhs*, the ones which match also the *rhs* are selected. Finally, a score is computed on the basis of this set of matching rules:

$$h_{\texttt{rules}}(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a}) = \frac{1}{K} \sum_{i=1}^{K} \log \Pr(r_i) \quad (1)$$

where r_i is a matching rule, $\Pr(r_i)$ its probability and K the number of the reordering patterns matching the given source/target pair.

2.3. Number of Source Phrases

During decoding, the source string is segmented into a sequence of phrases. Generally speaking, it is expected that the lower the number of source phrases covering the whole input string, the better the translation. In fact, if a single phrase from the phrase table is able to cover the input, likely the translation will be correct; on the contrary, a word-by-word translation is known to be worse than a phrase-based one. This fact suggests to use as additional feature the following score which combines the length of the input string and the number of phrases actually employed to cover it:

$$h_{PN}(f_1^J, e_1^I) = \frac{J}{N(\tilde{f})} \quad (2)$$

where $N(\tilde{f})$ is the number of source phrases.

3. Experiments

Experiments were carried out on the *Basic Traveling Expression Corpus* (BTEC) task [6]. BTEC is a multilingual speech corpus which contains sentences coming from phrase books for tourists traveling abroad. Training data, development sets

Table 3: Results of the optimization tecniques on the dev set 1 (BLEU% score and NIST score; without case nor punctuation).

System	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng		Ita-to-Eng	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
baseline	16.42	5.800	15.63	5.894	23.07	5.825	41.57	8.865
+CLA alignment	16.85	5.977	16.40	6.000	23.50	5.978	41.97	8.873
+Union IBM	16.99	6.171	17.01	6.087	23.90	6.023	–	–
+non-monotonic	18.87	6.388	19.57	6.828	24.35	6.112	–	–
+rescoring	21.82	6.793	22.96	6.989	25.88	6.160	44.54	9.110

Table 4: Results of the optimization tecniques on the dev set 2 (BLEU% score and NIST score; without case nor punctuation).

System	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng		Ita-to-Eng	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
baseline	14.84	5.295	13.38	5.324	19.21	5.276	35.24	7.779
+CLA alignment	15.03	5.475	13.94	5.475	19.59	5.301	35.59	7.859
+Union IBM	15.55	5.697	14.30	5.501	19.87	5.382	–	–
+non-monotonic	16.67	5.828	16.29	5.846	20.04	5.445	–	–
+rescoring	18.43	6.137	18.00	6.001	21.07	5.465	37.47	8.075

and test sets were provided. Statistics of the preprocessed corpora are shown in Table 1. We took part to the evaluations involving all the four language pairs of the Open Data track. It is worth noticing that the development sets were used to estimate the *weights* of the features/models, while the features/models themselves were estimated on the training sets. No additional resource was exploited for development/training purposes.

3.1. Preprocessing

The most important stage of preprocessing is tokenization. Actually, Chinese and Japanese are already tokenized in the supplied data; in order to smooth possible word segmentation inconsistencies, words of these two languages were re-segmented, as we did for Chinese in 2005 system. For other languages, the tokenization mainly separates punctuation from words. In addition, Arabic prefixed articles "*AL*" and "*BIL*" are also separated from words.

Preprocessing also includes the transformation of numbers written in textual form into digits and of upper case characters into lower case format. Table 2 shows which preprocessing stages are performed on different languages.

3.2. Postprocessing

IWSLT-06 evaluation is case sensitive and with punctuation. Since the source strings of test sets do not include punctuation marks, we developed two post-processing modules which work in cascade: first, the *punctuation restoration* module introduces the punctuation marks into the target string; second, the *case restoration* module recovers word case information of proper names, words after strong punctuation, etc.

For both the modules we use the `disambig` tool,[1] as suggested in the instructions supplied by the evaluation organizers. The English training data have been employed to train the language models of the two modules.

3.3. Development: Baseline and Improvements

The setup of baselines includes the use of phrases up to 8 words and monotonic search. We extracted phrases and estimated the phrase translation models from the intersection of direct and inverse IBM alignments, expanded as suggested in [7].

Improvements were carried out by introducing novelties in an incremental way, monitoring performance on development sets. Upgrades involve: (i) the addition of CLA word-alignments and (ii) of IBM union word-alignments [3], (iii) the execution of non-monotonic search, and (iv) the application of the rescoring module.

Non-monotonic search uses constraints on word reordering defined by means of the maximum vacancy number (MVN) and maximum vacancy distance (MVD) parameters. The following setting was used for experiments:

- Chinese-to-English: MVD=6 MVN=6
- Japanese-to-English: MVD=8 MVN=8
- Arabic-to-English: MVD=4 MVN=4
- Italian-to-English: MVD=0 MVN=0

Hence, concerning Italian-to-English pair, monotone search was performed since it resulted convenient in prelim-

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

Table 5: Contribution of each feature function in rescoring on dev set 1 (BLEU% and NIST scores; without case nor punctuation).

System	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng		Ita-to-Eng	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
baseline	18.87	6.388	19.57	6.828	24.35	6.112	41.97	8.873
- phrase number	21.47	6.724	22.57	6.950	25.50	6.162	44.41	9.118
- word reordering	21.32	6.727	22.44	6.974	25.41	6.157	44.46	9.106
- block reordering	21.28	6.714	22.30	6.966	25.37	6.167	44.53	9.104
- Dir. IBM Mod. 1	21.22	6.660	22.62	6.987	25.56	6.123	44.23	9.109
- Dir. IBM Mod. 3	21.34	6.669	22.69	6.985	25.60	6.117	44.17	9.112
- Inv. IBM Mod. 1	21.00	6.681	22.17	6.867	25.36	6.137	44.12	9.097
- Inv. IBM Mod. 3	21.16	6.727	22.58	6.653	25.42	6.175	44.24	9.103
- CLA score	21.18	6.600	22.59	6.936	25.76	6.032	44.22	9.041
- question feature	21.36	6.694	22.74	6.938	25.62	6.147	44.54	9.110
- n-gram frequency	20.50	6.458	21.81	6.400	25.03	5.441	43.21	8.667
- n-gram post-prob	21.23	6.716	22.89	6.974	25.70	6.164	44.28	9.123
- length prob.	21.52	6.692	22.64	6.950	25.47	5.949	44.35	8.959
- length ratio	21.15	6.710	22.74	6.982	25.60	6.080	44.39	9.084
- 2-grams LM	21.57	6.753	22.63	6.968	25.63	6.312	44.18	9.094
- 3-grams LM	20.95	6.693	22.00	6.776	25.23	6.164	43.63	9.106
- 5-grams LM	21.18	6.729	22.22	6.966	25.43	6.272	43.99	9.111
+ all features	21.82	6.793	22.96	6.989	25.88	6.160	44.54	9.110

inary experiments. It is worth to notice that monotone alignment of phrases does not prevent the rescoring module from reordering words. In fact, intra-phrase word reorderings can be observed and possibly awarded by reordering rules.

For rescoring, the set of additional feature functions listed in Section 2 was applied to lists of 5000-best translations.

Tables 3 and 4 show results measured on the two development sets; all the scores refer to case insensitive and without punctuation marks evaluation.

Table 5 and 6 show the contribution of each single additional feature functions used in re-scoring. Each entry of the table corresponds to the performance obtained by removing from the set of additional features just the feature under control. The weights of the remaining additional features are not re-estimated, i.e. their values are the same of those used to compute the scores of the last rows of tables ("+all features"). The gap between the entry and the value in the last row indicates how much important that feature is.

Table 7 shows results on the development sets as computed by the IWSLT 2006 submission server.

3.4. Performance Discussion

Figures of Tables 3 and 4 confirm what was shown in [3]: the use of alternative word-alignment models can improve the translation performance, especially for those language-pairs which have very different word order. By using both the CLA and IBM union word-alignments of training parallel texts, BLEU scores improved on the two development sets between 1% and 9% relative. Remarkably, for Japanese-to-English tasks, the absolute BLEU scores rose from 15.63 to 17.01 and from 13.38 to 14.30 on the two sets, respectively.

Non-monotonic search gave significant improvements in Chinese-to-English and Japanese-to-English tasks, since for these two language pairs the possibility of reordering words is necessary for covering their different grammatical structures. On the second development set, 7% relative BLEU score improvement was observed for Chinese-to-English (from 15.55 to 16.67), and 14% for Japanese-to-English (14.30 to 16.29). The translation from Arabic to English got some improvement from the non-monotonic search as well. Concerning Italian and English languages, since they have a similar word order, *phrases* seem sufficient to handle local reordering phenomena; this is the reason for which preliminary investigations suggested us not to relax the monotonic search constraint.

Rescoring yielded significant improvement of performance on development sets for all the language pairs. It is worth to highlight the contribution ensured by the additional features that are "new" with respect to the set employed in the 2005 evaluation.

Inverse IBM models had similar impact to that of direct IBM models.

Both word- and block-reordering rules improved the BLEU score of more than 2% relative for Chinese-to-English and Japanese-to-English, and of almost 2% relative for Arabic-to-English.

"n-gram frequency" feature function still plays a key role in our re-scoring models. Concerning the BLEU score, the relative improvements are from about 3% (Italian) to 6.4% (Chinese). In terms of NIST score, the relative improvements range from around 3% (Arabic) to 4.4% (Japanese).

Although there are no punctuation marks in the test sen-

Table 6: Contribution of each feature function in rescoring on dev set 2 (BLEU% and NIST scores; without case nor punctuation).

System	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng		Ita-to-Eng	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
baseline	16.67	5.828	16.29	5.846	20.04	5.445	35.59	7.859
- phrase number	18.23	6.099	17.87	5.993	20.86	5.457	37.16	8.015
- word reordering	17.99	6.101	17.48	6.018	20.68	5.464	37.42	8.074
- block reordering	17.85	6.105	17.38	5.949	20.61	5.456	37.39	8.071
- Dir. IBM Mod. 1	18.20	6.117	17.78	5.997	20.83	5.433	37.23	8.056
- Dir. IBM Mod. 3	18.25	6.058	17.85	5.997	20.84	5.443	37.18	7.987
- Inv. IBM Mod. 1	18.17	6.088	17.75	5.945	20.83	5.460	37.10	8.036
- Inv. IBM Mod. 3	18.22	6.110	17.75	5.989	20.81	5.455	37.24	8.071
- CLA score	18.19	6.093	17.82	6.000	20.76	5.359	37.15	8.017
- question feature	18.25	6.111	17.57	5.969	20.92	5.462	37.47	8.075
- n-gram frequency	17.65	5.882	17.01	5.748	20.39	5.304	36.33	7.806
- n-gram post-prob	18.19	6.117	17.75	5.993	20.88	5.475	36.95	8.051
- length prob.	18.27	6.064	17.68	5.980	20.98	5.469	37.35	7.943
- length ratio	18.22	6.109	17.85	5.991	20.92	5.408	37.08	8.000
- 2-grams LM	18.09	6.142	17.85	5.991	20.92	5.467	37.14	8.100
- 3-grams LM	17.60	6.064	17.18	5.980	20.53	5.476	36.79	8.066
- 5-grams LM	18.02	6.139	17.31	5.992	20.68	5.489	37.00	8.063
+ all features	18.43	6.137	18.00	6.001	21.07	5.465	37.47	8.075

Table 7: Scores on the two development sets (with case and punctuation).

Language	Dev. Set	BLEU%	NIST	WER%	PER%	METEOR%
Chi-to-Eng	Dev1 (Text)	20.56	6.225	67.81	52.57	50.22
	Dev2 (Read Speech)	17.74	5.635	71.33	56.99	45.94
Jpn-to-Eng	Dev1 (Text)	21.68	6.430	70.11	52.97	51.43
	Dev2 (Read Speech)	17.95	5.590	73.91	58.03	46.08
Ara-to-Eng	Dev1 (Text)	25.40	5.832	59.20	50.32	51.07
	Dev2 (Read Speech)	21.33	5.245	62.69	53.67	46.63
Ita-to-Eng	Dev1 (Text)	39.52	8.173	46.49	38.37	68.55
	Dev2 (Read Speech)	33.93	7.272	51.69	44.22	62.93

tences, the *question* feature still gave improvements on Chinese, Japanese and Arabic tasks. The possible reason could be that in those languages some words indicate the question form of the sentence: as an example, for a Chinese Yes-No question, the last word usually is "*ma*", which is typically aligned to the question mark of target sentence.

3.5. Performance on Test Sets

Table 8 shows the official scores on the test sets as reported by the submission server. In order to better understand the behavior of the system, Table 9 provides scores computed on the decoder output (1-pass) and after the rescoring step (2-pass), in case insensitive and without punctuation modality.

Finally, Table 10 compares the translation performance of 2005 and 2006 ITC-irst systems, measured on evaluation sets of 2004 and 2005 campaigns. Columns "2005" provide scores reported in [1]; columns "2006" contain BLEU scores computed by using the IWSLT 2006 evaluation server. Note that part of the improvement on Chinese-to-English and Japanese-to-English tasks comes from the larger amount of training data made available this year. Since the same training data were employed to develop the 2005 and 2006 systems for Arabic-to-English task, performance comparison on this language pair better assesses the improvement of our system.

Table 10: Comparison of the ITC-irst systems of the years 2005 and 2006 for the supplied data track (2005) and open data track (2006) on IWSLT04 and IWSLT05 test sets.

	Chi-to-Eng		Jpn-to-Eng		Ara-to-Eng	
BLEU%	2005	2006	2005	2006	2005	2006
IWSLT'04	46.37	53.23	50.11	54.01	56.37	58.14
IWSLT'05	52.75	59.91	43.13	54.83	56.22	57.57

Table 8: Official scores on the test sets (with case and punctuation).

Language	Test Set	BLEU%	NIST	WER%	PER%	METEOR%
Chi-to-Eng	Text	18.37	5.827	68.62	53.25	48.52
	Read Speech	15.60	5.221	72.01	57.86	43.74
	Spontaneous Speech	14.22	4.919	73.07	59.11	41.19
Jap-to-Eng	Text	18.39	5.854	71.60	54.23	47.44
	Read Speech	16.04	5.417	73.77	57.00	43.97
Ara-to-Eng	Text	20.05	5.182	63.22	53.89	45.81
	Read Speech	17.23	4.735	65.93	56.62	41.86
Ita-to-Eng	Text	34.97	7.816	48.22	39.98	64.68
	Read Speech	27.97	6.622	54.52	46.62	55.92

Table 9: Official scores of the 1-pass and 2-pass systems on the test sets (without case nor punctuation).

Language	Test Set	BLEU%		NIST		WER%		PER%		METEOR%	
		1-pass	2-pass	1-pass	2-pass	1-pass	2-pass	1-pass	2-pass	1-pass	2-pass
Chi-Eng	Text	16.94	19.92	6.043	6.426	74.01	70.52	55.25	51.59	46.10	48.52
	Read Speech	14.61	16.98	5.272	5.744	74.60	74.02	58.49	56.54	41.39	43.70
	Spont. Speech	13.44	15.77	4.961	5.480	75.84	75.49	59.80	57.82	38.90	41.21
Jpn-Eng	Text	16.61	18.82	6.010	6.320	74.59	73.42	54.94	53.07	45.54	48.24
	Read Speech	14.34	16.17	5.508	5.833	76.98	76.72	58.30	56.83	42.14	43.92
Ara-Eng	Text	19.75	20.48	5.524	5.604	64.69	64.01	53.75	53.02	44.93	45.64
	Read Speech	17.05	17.80	5.011	5.190	67.57	67.07	57.11	56.19	41.01	41.82
Ita-Eng	Text	37.10	37.97	8.489	8.619	45.90	45.60	35.45	35.27	63.79	64.61
	Read Speech	28.78	29.69	7.176	7.260	53.75	53.56	44.21	43.78	55.30	55.88

4. Conclusions

This work focused on novel aspects of the ITC-irst SMT system with respect to that employed in the evaluation campaign of the last year. Performance on development and test sets are reported in detail: this allows both to quantify the impact of techniques/modules on the quality of translations and to compare our 2006 system (i) with that we employed in 2005, and (ii) with those of other evaluation participants.

5. Acknowledgements

This work has been partly 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).

6. References

[1] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico, "The ITC-irst SMT System for IWSLT-2005," in *Proceedings of IWSLT*, 2005, http://www.is.cs.cmu.edu/iwslt2005/proceedings.html.

[2] M. Federico and N. Bertoldi, "A word-to-phrase statistical translation model," *ACM Transaction on Speech Language Processing*, vol. 2, no. 2, pp. 1–24, 2005.

[3] B. Chen and M. Federico, "Improving Phrase-Based Statistical Translation through Combination of Word Alignment," in *Proceedings of FinTAL - 5th International Conference on Natural Language Processing*, Turku, Finland, Aug. 2006, pp. 356–367.

[4] R. Zens and H. Ney, "N-Gram Posterior Probabilities for Statistical Machine Translation," in *Proceedings of the NAACL Workshop on SMT*, New York, USA, 2006.

[5] B. Chen, M. Cettolo, and M. Federico, "Reordering Rules for Phrase-based Statistical Machine Translation," in *Proceedings of IWSLT06*, www.slc.atr.jp/IWSLT2006.

[6] 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 *Proceedings of 3rd LREC*, Las Palmas, Spain, 2002, pp. 147–152.

[7] F. J. Och and H. Ney, "Improved Statistical Alignment Models," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, Hong Kong, China, 2000.