FBK @ IWSLT 2007

N. Bertoldi, M. Cettolo, R. Cattoni, M. Federico

FBK - Fondazione B. Kessler, Trento, Italy
surname@itc.it

Abstract

This paper reports on the participation of FBK (formerly ITC-irst) at the IWSLT 2007 Evaluation. FBK participated in three tasks, namely Chinese-to-English, Japanese-to-English, and Italian-to-English. With respect to last year, translation systems were developed with the `Moses` Toolkit and the `IRSTLM` library, both available as open source software. Moreover, several novel ideas were investigated: the use of confusion networks in input to manage ambiguity in punctuation, the estimation of an additional language model by means of the Google's Web 1T 5-gram collection, the combination of true case and lower case language models, and finally the use of multiple phrase-tables. By working on top of a state-of-the art baseline, experiments showed that the above methods accounted for significant BLEU score improvements.

1. Introduction

This paper presents work carried out at FBK (formerly ITC-irst) to develop speech translation systems for three translation tasks of the IWSLT 2007 Evaluation, namely Chinese-to-English (CE), Japanese-to-English (JE), and Italian-to-English (IE).

All three systems are based on different set-ups of the same translation engine, namely the `Moses` statistical machine translation (SMT) Toolkit [1].[1] The FBK team joined the `Moses` open source project in 2006 and indeed discontinued the development of its own decoding software [2].

This paper is organized as follows. Section 2 gives a short introduction of the general architecture of the systems, which basically takes advantage of experience gained in the past IWSLT evaluations. Section 3 focuses on the novel aspects that were investigated specifically for IWSLT 2007, namely the management of punctuation and the use of additional language resources. Sections 4 to 6 discuss details related to the development and experimentation of each single translation system. Section 7 summarizes set-ups and official results of the primary and contrastive submissions. Section 8 provides some general discussion and conclusions about our participation to IWSLT 2007.

[1] Available from http://www.statmt.org/moses

2. System description

In its simplest configuration, the speech translation system provides the translation of the best transcription hypothesis generated by the ASR module. This version is used in the manual transcription input condition. Nevertheless, a tighter ASR-SMT integration [3] is considered as well. Given the word lattice produced by any ASR engine: (i) the word lattice is transformed into a compact structure, called *confusion network* (CN); (ii) punctuation is inserted in the CN; (iii) the optimal translation is computed from the CN; (iv) finally, if needed, case information is added to the translation.

The insertion of punctuation (step ii), described in Section 3.1, is required whenever the final translation should be enriched with this information but it is not included in the input lattices.

2.1. Extraction of Confusion Network

A word lattice contains several transcription alternatives explored during the ASR process, but its topology is very complex and redundant. A simpler and more compact way of representing these alternatives is achieved through a CN [4]. A CN is still a weighted directed graph with the peculiarity that each path from the start node to the end node goes through all the other nodes; words and posterior probabilities are associated to the graph edges.
The extraction of a CN from a word lattice is done by means of the `lattice-tool` by SRILM toolkit [5].[2]

2.2. Decoder

The currently available release of `Moses` features a multi-stack, phrase-based, beam-search decoder able to process a CN as well as plain text. It implements a log-linear translation model including as feature functions: direct and inverse phrase-based and word-based lexicons, multiple word-based n-gram target language models (LMs), phrase and word penalties, and distance-based (possibly lexicalized) reordering model.
Computational efficiency is obtained through prefetching and early recombining the translation alternatives of the source phrases. On-demand loading of lexicon, distortion models and LMs [6] and quantization of LMs [7] allow a big reduction of run-time memory usage.

[2] Available from http://www.speech.sri.com/projects/srilm

A detailed description of the CN decoder can be found in [8].

2.3. Rescoring

After running `Moses`, a second decoding step can possibly be applied, which rescores the 5000-best list of translation hypotheses with the following 7 additional features:

- direct and inverse IBM Model 1 lexicon scores which should capture lexical co-occurrences in the source and target strings;
- weighted sum of n-grams relative frequencies (n from 1 to 4) in N-best list, which favors hypotheses containing popular n-grams of higher order;
- the reciprocal of the rank (log), which prefers high-ranked hypotheses;
- counts of hypothesis duplicates (log), which awards translations occurring several times;
- n-gram posterior probabilities within the N-best translations [9];
- sentence length posterior probabilities [9].

2.4. Capitalization

In the IE task both human and automatic transcripts do not contain case information. We decided to perform translation with models trained on lower-cased texts, and restore capitalization as postprocessing step, by means of the `disambig` tool of SRILM toolkit, fed with a n-gram case sensitive target LM.

Instead, in the CE and JE tasks this issue is not present because the source languages do not represent capitalization explicitly. Hence, case information is automatically introduced during translation by using case-sensitive models on the target language side.

2.5. Training

`Moses` toolkit also includes facilities to train the bilingual lexicons and the distortion models given a word-aligned parallel corpus, and to optimize feature weights on a development set through a Minimum Error Rate training (MERT).

In particular, phrase-based translation models (TMs) are estimated as follows. i) A parallel corpus is word-aligned by means of GIZA++ software tool [10] in both direct and inverse direction; ii) a list of phrase-pairs (up to 8 words) is extracted exploiting both word-alignments; iii) this collection is expanded with single word translations from direct alignment; iv) each phrase pair is associated with direct and inverse phrase-based and word-based probabilities.

The MERT procedure implemented by `Moses` toolkit applies an iterative and (locally) convergent strategy to find a set of weights which optimizes the BLEU score; a 5000-best list of translations provided by the decoder is exploited for this purpose after each translation step, and possibly after the rescoring step.

`Moses` is also able to train, load and exploit very huge LMs, through the `IRSTLM` library, an open source software developed at IRST [6].[3] Word-based LMs are trained with modified Kneser-Ney discounting.

3. Novelties

3.1. Punctuating Confusion Networks

An issue of SLT is the management of punctuation. At present, most ASR systems do not generate transcription hypotheses (N-best or word graphs) including punctuation information. Nevertheless, final translations are required to include punctuation to improve readability and comprehension.

Recent experiments on a large-vocabulary speech translation task from English-to-Spanish [11] have shown that adding punctuation to the ASR output is more effective than inserting it as a postprocessing step after translation. A further advantage of this approach is that it allows to exploit translation models trained on punctuated texts, both to translate written text and ASR output.

The method is now briefly described assuming that the ASR output is a CN, but it trivially applies to text input as well (see [11] for a comprehensive description).

i. The consensus decoding [4] is extracted from the CN;

ii. multiple hypotheses of punctuation marks are generated by means of the `hidden-ngram` tool by SRILM toolkit; it is worth noticing here that in this step a limited set of punctuation marks (namely strong marks only) is used;

iii. a new CN with strong punctuation marks is created from the multiple hypotheses provided by the tool;

iv. this punctuated CN is merged with the original CN;

v. Steps i) to iv) are re-executed using an extended set of punctuation marks that includes weak marks like commas, colon, etc.

At the end, the method provides a CN that represents the alternatives both on words, originally provided by ASR, and on punctuation marks. The procedure is a clear improvement of [12], as punctuation hypotheses are inserted in the CN through a statistical model and not deterministically.

Insertion of punctuation was applied to the IE task only, where both human and ASR transcripts did not contain such information. Punctuation insertion was not applied in the ASR condition of the CE and JE tasks, mainly due to our poor knowledge of Chinese and Japanese and lack of time to carry out a reasonable investigation. In conclusion, we handled ASR output as it were text.

[3] Available from http://sourceforge.net/projects/irstlm

5-gram	*fr*	$\lfloor \log_{10}(fr) \rfloor$	*expansion*
a b c d e	42	1	a b c d e ...
a b c d a	122	2	a b c d a ... a b c d a ...
a b c d b	3892	3	a b c d b ... a b c d b ... a b c d b ...

Table 1: Expansion of three 5-grams from the Google Web 5-grams. The final corpus is obtained by concatenating all 5-gram expansions.

3.2. LM with the Google Web n-grams

One of the resources we considered to overcome the limited availability of monolingual data in the target language was the so called *Web 1T 5-gram version 1* collection, distributed by the Linguistic Data Consortium. This data set, developed by Google Inc., contains English word n-grams, up to to 5-grams, and their observed frequency counts. n-grams were extracted from approximately 1 trillion word tokens of text from publicly accessible Web pages.

Indeed, the collection contains a fraction of the actually observed n-grams, namely only those with frequency count of at least 40. Moreover, word tokens occurring less than 200 times were all replaced with the fixed token `<UNK>`. This filtering results in a significantly lower but still respectable number of statistics, namely 1.2 billion 5-grams over a vocabulary of 13.5 million words.

The Google Web n-gram collection poses problems both from the modeling and implementation point of view. With respect to modeling, we decided to use the Google Web 5-grams as a source to augment the 5-gram statistics of the IWSLT training data. In order to focus on domain related events, we selected Google 5-grams such that all their 1-grams and a given percentage of their 2-grams are contained in the training data. In particular, tests were conducted by assuming two different percentages on the 2-grams, namely 75% and 100%: that is, in one case a Google 5-gram is chosen only if all its 1-grams and at least 3 out of 4 of its 2-grams are observed in the training data; in the other case, only if all its 1- and 2-grams occur in the training data. Once Google 5-grams were extracted, a corpus was artificially created in order to estimate a LM with consistent statistics for all n-gram levels. The corpus was generated by concatenating each selected Google 5-gram a number of times equal to the $\log_{10}$ of its original frequency count. The conventional word "..." is interleaved to avoid the generation of spurious 5-grams. Table 1 sketches the procedure.

The resulting corpus was used to estimate a LM using the improved Kneser-Ney smoothing method. Notice that the expansion naturally guarantees that for all 5-grams in the corpus, each suffix is contained in the prefix of at least another 5-gram. This property is necessary to build up a consistent LM table. Notice that the table will also contain many 5-grams with the spurious concatenation symbol. However, such entries are necessary to access lower-order n-grams.

3.3. Case Sensitive Versus Case Insensitive LMs

Typically, in SMT systems the dictionary of the LMs coincides with that of the target side of the translation and reordering models. When the latter is case sensitive, the importance of having a case sensitive LM is even higher, because it happens that the decision whether capitalizing or not a given word can be taken only by looking at its context, being the phrase tables unable to provide the proper hint.

Of course, case information found in LM training data should be coherent with that of the other models. Since LMs estimated on huge corpora can significantly help translation performance, the problem of capitalization consistency among different language resources arises.

A possible solution is to train case sensitive models only from reliable (with respect to the task) sources and to ignore case information from other additional corpora. To this aim, we introduced a general extension to `Moses` which permits to introduce a word-to-word map between the target language and the dictionary of the LM. In this way, a case insensitive LM can be queried with case sensitive n-grams through a TrueCase-NoCase word map.

3.4. Handling of Huge LMs

With respect to the past IWSLT evaluations, the possibility of exploiting additional language resources has brought up the need to efficiently handle huge LMs at decoding time. For this purpose, we have used a version of `Moses` compiled with the `IRSTLM` library [6]. This library provides efficient data structures that permit to store and access huge n-gram LMs with a reasonable trade-off between space and time requirements.

3.5. Multiple Phrase Tables

An under-investigated ingredient of `Moses` is the possibility of exploiting multiple phrase tables at the same time. The decoder allows two possibilities: (i) each partial translation is scored by interpolating the scores from all phrase tables (in this case, only phrase pairs belonging to the intersection of phrase tables can be scored), or (ii) a partial translation is activated for each phrase table. We worked on the latter case, leaving the experimental comparison of the two alternatives for future investigations. Then, in decoding stage of our experiments, `Moses` accesses a phrase table to get translation options ranked according to four different scores (if the default setting is employed). Concerning the collection of translation options, there is no difference between using one phrase table or multiple phrase tables trained on a partition of the original training data. There is however an important difference from the scoring point of view. Different scores are estimated and different weights in the log-linear model can be assigned to each table. This permits to better manage translation options coming from non homogeneous training sets and to let the MERT algorithm learn their respective utility inside the log-linear model.

However, a drawback of using multiple phrase tables is the increase in complexity introduced in the log-linear model, which reflects negatively on search time and on the weight optimization procedure.

4. Chinese-to-English System

FBK developed a system for the CE challenge translation task, namely the translation of spontaneous conversations in the travel domain. Although actual evaluation data consisted in transcriptions of read speech (classical task), we did not re-tune the system to the new type of data and kept the best setup defined by the following development experiments.

4.1. Data

Table 3 gives statistics of corpora employed for the CE task. Figures refer to texts after preprocessing which mainly consists in re-segmentation of Chinese words, as we did in previous IWSLT evaluations, and English tokenization. Case information was kept; digits were converted into textual form.

For training, in addition to the Basic Travel Expression Corpus (`BTEC`) data we also used the evaluation sets of 2003 (`dev1`), 2004 (`dev2`) and 2005 (`dev3`), all of which can be considered belonging to the classical task. On the contrary, the development and test sets of the last evaluation campaign relate to the challenge task and for this reason we decided to tune the system on them. In particular, the development set (`dev4`) was first used for the MERT-based estimation of the interpolation parameters, and then added to the training data for the estimation of the models. Two versions of the test set were employed to assess the quality of the resulting system, namely the verbatim transcription (`dev5-vrb`) and the 1-best output of an automatic transcription (`dev5-1bst`).

We also exploited parallel corpora from the LDC and within the list provided by the organizers of IWSLT (`LDC`).

set	usage	source		target	
		|W|	|V|	|W|	|V|
`BTEC`	train	351,588	11,343	365,097	11,320
`LDC`	train	83.9M	277K	90.4M	342K
`dev123`	train	10,957	1,898	196,804⋆	4,660⋆
`dev4`	dev	5,137	1,174	45,720⋆	2,150⋆
`dev5-vrb`	test	5,599	1,350	51,227⋆	2,346⋆
`dev5-1bst`	test	5,487	1,380	”	”

Table 3: CE task: statistics of training, development and test sets. (⋆) These figures refer to 7/16 references.

4.2. Baseline Setup

Most of the features of the CE baseline system are provided in Section 2 and Table 2. The corpus `BTEC+dev1234` was employed for the estimation of the LM, the TM and the case sensitive “orientation-bidirectional-fe” reordering model [13]. For what is not specified, `Moses` default settings were employed.

4.3. Results

Translation results of both human transcripts (`vrb`) and 1-best ASR (`1bst`) of `dev5` are given in Table 4. The first row reports performance of the baseline system as defined in the previous section. Values refer to BLEU% and NIST scores.

in	TM_1	TM_2	LM_2	BLEU%	NIST
`vrb`	`baseline`	-	-	20.50	5.57
	`nc align`	-	-	21.86	5.59
	`+union`	-	-	22.35	6.20
	`+inter.`	-	-	22.71	6.31
	”	-	`web100`	22.00	6.25
	”	-	`web75`	22.54	6.30
	”	-	`web75nc`	22.85	6.31
	”	`LDC`	`web75nc`	**23.50**	**6.62**
`1bst`	”	”	”	**18.64**	**5.66**

Table 4: CE task: case sensitive with punctuation scores on 2006 evaluation set.

To reduce data sparseness in the BTEC training data, we removed case information before estimating word alignments and restored it in the successive steps, namely phrase extraction and estimation of translation and reordering models. The rows `baseline` and `nc align` (no case alignment) in Table 4 show the impact of this preprocessing step on the performance of a baseline.

After that, we enlarged the TM with phrases extracted by applying alternative symmetric alignment methods [14], namely the union and intersection heuristic. Performance are shown in `+union` and `+inter.` rows, respectively. Both these extensions yield a performance increase of BLEU and NIST scores with respect to the baseline employing the so called grow-diagonal heuristics [10].

With the just detailed TM, we tried to use a second LM in addition to the 6-gram baseline LM. We exploited the Google Web n-grams as described in Section 3.2. In particular, first, we integrated a case sensitive Web LM of 101M 5-grams consisting of 2-grams occurring in the `BTEC+dev123` training data (entry `web100` in Table 4). Then, other 253M 5-grams containing just 3 (out of 4) 2-grams occurring in training data were included (entry `web75`). Finally, since we noted significant discrepancies on case information between Web n-grams and BTEC texts, we exploited the `Moses` upgrade presented in Section 3.3 for integrating a case insensitive Web LM (336M 5-grams) into our case sensitive system (entry `web75nc`). It can be noted that only in the latter case the Web LM helps a bit; moreover, it is more effective to add its case insensitive version than the case sensitive one. Anyway, further investigations are needed for effectively exploiting Web n-grams in a task-dependent system.

Finally, we added a second TM (`LDC`) trained on case sensitive data from LDC. In this case, texts were not lowered before computing alignments, and phrases included up to 6

task	LM		TM			decoding		
	case sensitive	size	case sensitive	texts for word alignments	heuristics for word alignments	stack size	phrase table limit	reordering limit
CE	Y	6	Y	true case	grow-diag-final	2000	50	7
JE	Y	6	Y	no case	grow-diag-final + union + inters.	2000	50	7
IE	N	5	N	no case	grow-diag-final + union	200	20	6

Table 2: CE, JE and IE baseline setups.

set	usage	source		target	
		|W|	|V|	|W|	|V|
BTEC	train	401,628	12,428	365,116	11,320
Reuters	train	1.68M	63.0K	1.52M	35.3K
dev123	train	12,472	1,872	196,804⋆	4,660⋆
dev4	dev	5,852	1,196	45,720⋆	2,150⋆
dev5-vrb	test	6,489	1,331	51,227⋆	2,346⋆
dev5-1bst	test	6,638	1,388	”	”

Table 5: JE task: statistics of training, development and test sets. (⋆) These figures refer to 7/16 references.

words. Performance of this 2-TMs/2-LMs system are given in the last but one row of the table and are clearly better than those of the 1-TM system. The scoring of translations of the recognizer output by this system is provided in the last row of Table 4: as for correct transcriptions, also in this case performance well compare with the best official results of the 2006 IWSLT evaluation campaign.

5. Japanese-to-English System

FBK also developed a system for the JE classic translation task, that is the translation of read speech in the travel domain. We exploited most of the outcomes of the CE system development, with few adjustments as specified in the following.

Statistics of corpora utilized here are given in Table 5. It can be noted the symmetry with respect to the resources used for the CE task, with the LDC parallel corpus replaced by the much smaller Reuters corpus [15], one of the shared resources.[4]

Performance of the various system configurations tested for the JE task are collected in Table 6. The first row refers to the baseline system (Section 2 and Table 2), that is the 1-TM/1-LM system corresponding to the entry +inter. of Table 4.

In this task, the attempt of adding the 5-gram web75nc LM did not give any improvement (second row). On the contrary, the use of a second TM trained on the Reuters corpus yielded to the best performing system (third row), which was also tested on the best automatic transcription of the Japanese speech (last row). Even for this task, our best performance

[4]Actually, although the Reuters corpus is provided with alignments, we re-aligned it by means of the GIZA++ tool [10].

in	TM_2	LM_2	BLEU%	NIST
vrb	-	-	22.98	6.09
	-	web75nc	22.89	5.94
	Reuters	-	**23.40**	**6.14**
1bst	Reuters	-	**20.89**	**5.71**

Table 6: JE task: case sensitive with punctuation scores on 2006 evaluation set.

well compare with the top scores of the 2006 evaluation.

6. Italian-to-English System

FBK participated to the Italian-to-English (IE) challenge task, namely the translation of spontaneous conversations in the travel domain. More specifically, the development and test data are translations of spontaneous speech extracted from the ADAM corpus developed under the SITAL project [16], a speech corpus containing dialogs between a travel agency's operator and a person seeking travel information or to book a ticket, a hotel room or a flight.
Two conditions are considered for the IE challenge task according to the type of input: (i) manual transcripts of the dialogues; (ii) output of an ASR system. In the latter case, the best automatic transcription, the list of the 20 best automatic transcriptions or the whole lattice are available. Automatic transcriptions were provided by FBK.

6.1. Data

The core set of the training data for this task consists of the Italian-English parallel corpus extracted from the BTEC, although the domain of the task is highly different. Among the publicly available linguistic resources allowed for the task, we took into account the Italian-English EuroParl corpus (EP), the Italian-English dictionary extracted from MultiWordNet[5] (MWN), and the Google Web 1T 5-grams (web). A set of 466M 5-grams was obtained after the application of the filtering procedure described in Section 3.2: all 5-grams containing at least 75% (3 out of 4) 2-grams occurring in training data and development sets were included.

Furthermore, a list of 276 Italian Named Entities (NE) from the NESPOLE![6] and SITAL projects are available; as their correct English translations were not provided, blind

[5]Available from http://multiwordnet.itc.it
[6]http://nespole.itc.it

verbatim translations are taken into account.

Table 7 reports statistics (number of running words and vocabulary size) of corpora employed for the IE task. Figures refer to preprocessed texts, which were tokenized and put in lower-case letter. Figures related to the source part of dev* refer to the human transcripts; figures related to the target part refer to all available references.

The development sets (dev*) were used for either training, development or test purposes. In the former case, a parallel corpus was obtained by pairing the source text with each available reference. 16 references are available for d1, d2, and d3, 7 references for d4 and d5a, and only one reference for d5b*.

The IE challenge task of IWSLT 2007 is focused on the SITAL domain and significantly differs from previous years' editions, which were strictly based on the BTEC domain. Accordingly, d1-d4 and d5a, which were built for the previous evaluations, do not properly fit the new task domain. Henceforth, a new task-related development set d5b was provided mainly for task-adaptation purposes. Moreover, rich lattices were provided only for d5b, which are comparable with those distributed as test data. We divided this new development set in two parts (d5b1 and d5b2), preserving dialogues' integrity, to have a better tuning of the system.

Table 8 reports the Graph Error Rates (GERs) of for d5b* and test for different kinds of speech input.

set	usage	source		target	
		|W|	|V|	|W|	|V|
BTEC	train	172K	10,160	183K	7,298
MWN	train	164K	38,954	170K	38,918
NE	train	495	260	495	260
EP	train	28M	115K	29M	83K
d1	train	3,464	1,067	65,622	2,139
d2	train	3,404	1,085	64,896	2,238
d3	dev/train	3,489	1,095	66,286	2,328
d4	train	4,831	1,233	45,720	1,823
d5a	test/train	5,607	1,467	51,227	2,042
d5b	train	8,485	868	11,730	723
- d5b1	dev/test/train	4,237	569	5,844	518
- d5b2	dev/test/train	4,248	696	5,886	593
test	test	6,421	735	-	-

Table 7: IE task: statistics of training, development and test sets (number of running words, size of the dictionary for source and target, usage).

6.2. Baseline Setup

Most of the features of the IE system are provided in Section 2 and Table 2. For what is not specified, Moses default settings were employed.

CNs were pruned through the removal of columns whose empty words have probability larger than 0.99. This pruning significantly reduces CN size, and only slightly affects Graph Error Rate, as shown in Table 8.

set	1-best	lattice	cn
d5b	10.29	3.73	3.91
- d5b1	8.98	3.27	3.43
- d5b2	11.58	4.18	4.39
test	10.70	4.39	4.41

Table 8: IE task: Graph Error Rates of the dev. and test sets. Figures refer to the 1-best ASR, to lattices, and to CNs.

Punctuation is added as explained in Section 3.1 to both text and confusion network inputs. A 10M 3-grams lowercase LM estimated on the Italian part of the BTEC and EP was exploited to feed hidden-ngram.

6.3. Results

In order to establish whether and which additional resources could increase performance, we first computed the Out-of-Vocabulary (OOV) rate of the two development sets d5b1 and d5b2 when different combination of training data are taken into account (see Table 9). The baseline OOV rate confirms that the new task is quite different from the BTEC domain. An absolute OOV rate reduction of 3.3-3.5 is obtained by adding the task-related IE NE list. The effect of BTEC-domain development sets (d1-d5a) is rather low. The addition of task-specific development sets (d5b*) is definitely important for adaptation; in fact, they reduce OOV rate by at least 1.5 absolute. The OOV rate reduction given by EP is mainly due to the large size of its vocabulary.

	OOV rate	
	d5b1	d5b2
BTEC	6.49	7.10
BTEC, NE	2.97	3.83
BTEC, NE, MWN (=BNM)	2.53	3.24
BNM, d1-d5a	2.27	3.17
BNM, d1-d5a, d5b1	-	1.68
BNM, d1-d5a, d5b2	0.77	-
BNM, EP	0.72	0.55
BNM, EP , d1-d5a	0.72	0.55
BNM, EP , d1-d5a, d5b1	-	0.28
BNM, EP , d1-d5a, d5b2	0.28	-

Table 9: IE task: OOV rate of the human transcripts of development sets d5b1 and d5b2 with different training corpora.

According to these preliminary results we decided to train the models of the baseline system on BTEC, NE, and MWN (BNM). In the following experiments, we analyzed the effect of using EP and development sets d* from the point of view of the translation quality.

A preliminary experiment was performed to validate the quality of the policy for inserting punctuation described in Section 3.1. We translated the human transcripts (vrb) of

d5b2 enriched with either the best punctuation hypothesis (vrb-bst) or the multiple punctuation hypotheses in the form of CN (vrb-cn). As reported in Table 10, the latter approach is definitely better than the former, and it was applied to all further experiments when input requires the insertion of punctuation. The absolute gain is larger than 2 BLEU points.

We also exploited additional resources to establish how much they impact over performance. Experiments, whose results are given in Table 10, were performed on CN-repunctuated human transcripts (vrb-cn) of d5b2; weights were optimized over d5b1.

It is worth noticing that the larger improvement is due to the exploitation of the task-related development set d5b1, which boosts BLEU score by 38.7% (27.50 vs 19.83). The use of the other development sets (d1-d5a), EP and web gives further improvement.

We preferred to add EP training data as additional TM and LM for two reasons. First, the EP domain is very far from the task domain, and hence separate weights could better balance their contribution. Secondly, we could estimate the TM and LM on EP independently (and once) and just add them to the decoder when needed.

input	TM_1,LM_1	TM_2,LM_2	LM_3	BLEU	NIST
vrb-bst	BNM	-	-	17.71	4.90
vrb-cn	BNM	-	-	19.83	4.95
	”	d5b1	-	27.50	5.66
	+d1-d5b1	-	-	28.70	5.76
	”	-	web	29.66	5.83
	”	EP	”	30.79	5.92

Table 10: IE task: case-insensitive with punctuation scores on d5b2.

In order to exploit the development set closer to the evaluation conditions (d5b) both in model training and in weights estimation, it was split into two parts (d5b1, d5b2) which were used as follows. First, we optimized the decoder weights on d5b1 using d5b2 as training data and got the weights wg1; then, we inverted the sets and started the optimization procedure from wg1 getting the final optimized weights wg2. As shown in Table 11, performance achieved with wg1 and wg2 are very close on the average, but the second one gives more balanced results.

The rescoring step described in Section 2.3 gives a further small improvement, as reported in Table 11.

At the end the module for case restoring is applied, which was fed with a 4-gram case-sensitive LM trained on all English training data. The case-sensitive evaluation shows a performance decrement not larger than 5.5%, as shown in **bold** in Table 11.

Tuning of the system was performed separately for the two kinds of input: namely, the human transcriptions and the CNs. Performance achieved by the system on the repunctuated CNs (cn) are reported in Table 11. A relative BLEU score decrement of 10.8% and 5.8% was observed on d5b1 and d5b2, respectively.

The former system was also applied to run contrastive experiments on the CN-repunctuated 1-best ASR (1bst-cn).

input		d5b1		d5b2	
		BLEU	NIST	BLEU	NIST
vrb-cn	wg1	33.89	6.22	30.79	5.92
	wg2	33.27	6.15	31.56	5.94
	+resc	33.20	6.19	32.06	6.00
	+case	**31.43**	**5.94**	**30.29**	**5.75**
cn	”	**28.04**	**5.68**	**28.54**	**5.55**

Table 11: IE task: Case-insensitive with punctuation scores on the CN-repunctuated human transcripts of d5b1 and d5b2. Figures in **bold** refer to case-sensitive evaluation.

7. Official Runs

Sections 4 to 6 presented the development of the systems employed in the three tasks of IWSLT 2007 Evaluation Campaign, which FBK participated to. To train the final systems we exploited the best set-ups, whose results are highlighted (in **bold**) in Tables 4, 6 and 11, and all development sets (in particular, dev5 for CE and JE, and d5b* for IE) as training data. Statistics about the TMs and LMs of final systems are reported in Table 12.

Official BLEU scores achieved by primary and some contrastive submissions are summarized in Table 13.

8. Conclusion

By considering the preparatory work and the overall results by FBK (formerly ITC-irst) in the IWSLT 2007 evaluation, the following conclusive remarks can be stated:

- Moses provides an excellent baseline to develop state-of-the-art SMT systems, on top of which further improvements and extensions can be integrated and made available to the research community.
- Confusion network decoding shows to be a statistically sound and effective way to manage ambiguous inputs, such as alternative ASR and punctuation hypotheses.
- The use of additional monolingual resources, such as the Web 1T 5-grams, has shown to be beneficial at least for some translation directions. However, improvements have been obtained by properly handling case information and by tuning the log-linear model so that proper relative weights are estimated for in-domain and out-of-domain data.
- The combination of parallel resources from heterogeneous sources can be effectively handled by generat-

task	TM_1	TM_2	LM_1		LM_2		LM_3	
			n	$\|n$-gr$\|$	n	$\|n$-gr$\|$	n	$\|n$-gr$\|$
CE	5.9M	27M	6	39K	5	336M	-	-
JE	9.1M	176K	6	39K	-	-	-	-
IE	3.8M	39M	4	362K	4	16M	5	466M

Table 12: Statistics of final systems models: number of phrase pairs of TMs, order and number of n-grams of LMs.

task	input	set-up	BLEU
CE, clean	01, `vrb`		**34.72**
	05, `vrb`	no TM_2, LM_2	35.08
JE, clean	01, `vrb`		**47.89**
	02, `vrb`	no `dev5`	48.93
JE, ASR	01, `1bst`		**39.46**
	04, `cn`		39.69
IE, clean	01, `vrb-cn`		**44.32**
	03, `vrb-cn`	no TM_2, $LM_{2/3}$	43.41
IE, ASR	01, `cn`		**42.29**
	05, `1bst-cn`		41.51

Table 13: BLEU scores of primary and (some) contrastive submissions: differences from primary set-up are reported. Figures in **bold** refer to primary runs.

ing translation hypotheses through alternative phrase-tables, which are estimated independently and properly weighted in the log-linear model.

IWSLT provides and excellent benchmark to evaluate novel ideas in the area of speech translation. Future work will however verify how our findings apply to larger and more complex translation tasks, such as those proposed by the TC-STAR and NIST evaluations.

9. References

[1] 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 *Proc. of ACL*, Prague, Czech Republic, 2007.

[2] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico, "The ITC-irst SMT System for IWSLT-2006," in *Proc. of IWSLT*. Kyoto, Japan, November 2006.

[3] D. Falavigna, N. Bertoldi, F. Brugnara, R. Cattoni, M. Cettolo, B. Chen, M. Federico, D. Giuliani, R. Gretter, D. Gupta, and D. Seppi, "The IRST English-Spanish Translation System for European Parliament Speeches," in *Proc. of Interspeech*, Antwerp, Belgium, August 2007.

[4] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[5] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of ICSLP*, Denver, Colorado, 2002.

[6] M. Federico and M. Cettolo, "Efficient Handling of N-gram Language Models for Statistical Machine Translation," in *Proc. of ACL-SMT workshop*, Prague, Czech Republic, 2007.

[7] M. Federico and N. Bertoldi, "How Many Bits Are Needed To Store Probabilities for Phrase-Based Translation?" in *Proc. on the ACL Workshop on Statistical Machine Translation*, New York, NJ, USA, June 2006, pp. 94–101.

[8] N. Bertoldi, R. Zens, and M. Federico, "Speech Translation by Confusion Network Decoding," in *Proc. of ICASSP*, Honolulu, Hawaii, USA, April 2007.

[9] R. Zens and H. Ney, "N-gram Posterior Probabilities for Statistical Machine Translation," in *Proc. on the ACL Workshop on Statistical Machine Translation*, New York, NJ, USA, June 2006, pp. 72–77.

[10] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[11] R. Cattoni, N. Bertoldi, and M. Federico, "Punctuating Confusion Networks for Speech Translation," in *Proc. of the Interspeech-2007*, Antwerp, Belgium, August 2007.

[12] W. Shen, R. Zens, N. Bertoldi, and M. Federico, "The JHU Workshop 2006 IWSLT System," in *Proc. of IWSLT*, Kyoto, Japan, 2006.

[13] 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 *Proc. of IWSLT*, Pittsburgh, PA, 2005.

[14] B. Chen and M. Federico, "Improving Phrase-Based Statistical Translation through Combination of Word Alignment," in *FinTAL*. Turku, Finland: Springer Verlag, LNCS, August 2006, pp. 356–367.

[15] M. Utiyama and H. Isahara, "Reliable Measures for Aligning Japanese-English News Articles and Sentences," in *Proc. of ACL*, Sapporo, Japan, 2003.

[16] R. Cattoni, M. Danieli, A. Panizza, V. Sandrini, and C. Soria, "ADAM: The SI-TAL Corpus of Annotated Dialogues," in *Proc. of LREC 2002*, Las Palmas, Canary Islands, Spain, 29-31, May 2002.