FBK @ IWSLT 2011

N. Ruiz [*], A. Bisazza, F. Brugnara, D. Falavigna, D. Giuliani, S. Jaber, R. Gretter, M. Federico

Fondazione Bruno Kessler-IRST
Via Sommarive 18, 38123 Povo (TN), Italy
nicruiz@fbk.eu

Abstract

This paper reports on the participation of FBK at the IWSLT 2011 Evaluation: namely in the English ASR track, the Arabic-English MT track and the English-French MT and SLT tracks. Our ASR system features acoustic models trained on a portion of the TED talk recordings that was automatically selected according to the fidelity of the provided transcriptions. Three decoding steps are performed interleaved by acoustic feature normalization and acoustic model adaptation. Concerning the MT and SLT systems, besides language specific pre-processing and the automatic introduction of punctuation in the ASR output, two major improvements are reported over our last year baselines. First, we applied a fill-up method for phrase-table adaptation; second, we explored the use of hybrid class-based language models to better capture the language style of public speeches.

1. Introduction

The IWSLT 2011 Evaluation Campaign [1] focused on the translation of TED Talks [1]: a collection of public speeches on a variety of topics. The evaluation campaign encompassed several tracks: (1) an ASR track, in which automatic transcriptions of talks were generated from audio (in English); (2) a SLT track, consisting of speech translation of talks from audio (or ASR output) to text for the English-French language pair; (3) a MT track, consisting of the translation of TED transcripts from English-French, Arabic-English, and Chinese-English; and (4) a SC track, which utilizes system combinations of ASR outputs (in English) and MT outputs (in English and French).
FBK participated in all evaluation tracks. This paper describes the automatic transcription system for English used in the ASR track and the Arabic-English and English-French MT systems used in the MT and SLT tracks.

2. ASR Task

In this section we summarize the main features of the FBK primary system used in the IWSLT 2011 Evaluation Campaign for transcribing TED talks delivered in English. For each talk, in addition to the audio file, time boundaries of speech segments to be transcribed are given. The word transcription of a talk is generated in three decoding passes. All the decoding passes make use of a 4-gram language model and are interleaved by acoustic feature normalization and acoustic model (AM) adaptation.

2.1. Acoustic data selection for training

For AM training, domain specific acoustic data were exploited. Recordings of TED talks released before the cut-off date, 31 December 2010, were downloaded with the corresponding subtitles which are content-only transcriptions of the speech. In content-only transcriptions anything irrelevant to the content is ignored, including most non-verbal sounds, false starts, repetitions, incomplete or revised sentences and superfluous speech by the speaker. A simple but robust procedure was implemented to select only audio data with an accurate transcription.

The collected data consisted in 820 talks, for a total duration of ∼216 hours, with ∼166 hours of actual speech. The provided subtitles are not a verbatim transcription of the speeches, hence the following procedure was applied to extract segments that can be deemed reliable. The approach is that of selecting only those portions in which the human transcription and an automatic transcription agree. To this end, a "background" 4-gram language model was first trained on all the talk transcriptions. Subsequently, a specific language model (LM) was built for each talk by adapting the language model to the human transcription of the talk. A preliminary automatic transcription was performed on the talks with a pre-trained general AM for English and the talk-specific LM. The output of the system was aligned with the reference transcriptions, and the matching segments were selected, resulting in an overlap of ∼120 hours of actual speech out of the total of 166. By using these segments together with the segments labeled as silence, a TED-specific acoustic model was trained, as detailed in the following section. The label/select/train procedure was repeated two more times, resulting in a portion of selected actual speech that grew to ∼142 hours and then to ∼144 hours. Given the modest improvement in the third iteration, the procedure was not repeated further. In conclusion, the method made available 87% of the training speech, which was considered satisfactory.

[*] The author was supported by the University of Trento.

[1] http://www.ted.com/talks

2.2. Acoustic model

Thirteen Mel-frequency cepstral coefficients, including the zero order coefficient, are computed every 10ms using a Hamming window of 20ms length. First, second and third order time derivatives are computed after segment-based cepstral mean subtraction to form 52-dimensional feature vectors. Acoustic features are normalized and HLDA-projected to obtain 39-dimensional feature vectors as described below.

AMs were trained exploiting a variant of the speaker adaptive training method based on Constrained Maximum Likelihood Linear Regression (CMLLR) [2]. In our training variant [3, 4, 5] there are two sets of AMs: the target models and the recognition models. For each cluster of speech segments, an affine transformation is estimated through CMLLR [2] with the aim of minimizing the mismatch between the cluster data and the target models. Once estimated, the affine transformation is applied to cluster data in order to normalize acoustic features with respect to the target models. Recognition models are then trained on the normalized data. Leveraging on the possibility that the structure of the target and recognition models can be determined independently, a Gaussian Mixture Model (GMM) can be adopted as the target model for training AMs used in the first decoding pass [3]. This has the advantage that, at recognition time, word transcriptions of test utterances are not required for estimating feature transformations. Instead, target models for training recognition models used in a second or third decoding pass are usually triphones with a single Gaussian per state [4]. In all cases, the same target models are used for estimating cluster-specific transformations during training and recognition.

In the current version of the system, a projection of the acoustic feature space based on Heteroscedastic Linear Discriminant Analysis (HLDA) is embedded in the feature extraction process as follows. A GMM with 1024 Gaussian components is first trained on an extended acoustic feature set consisting of static acoustic features plus their first, second and third order time derivatives. For each cluster of speech segments, a CMLLR transformation is then estimated w.r.t. the GMM and applied to acoustic observations. After normalizing the training data, an HLDA transformation is estimated w.r.t. a set of state-tied, cross-word, gender-independent triphone Hidden Markov Models (HMMs) with a single Gaussian per state, trained on the extended set of normalized features. The HLDA transformation is then applied to project the extended set of normalized features in a lower dimensional feature space, that is a 39-dimensional feature space. Recognition models used in the first and subsequent decoding passes are trained from scratch on normalized HLDA-projected features. HMMs for the first decoding pass are trained through a conventional maximum likelihood procedure. Recognition models used in the second or third decoding pass are speaker-adaptively trained, exploiting as target-models triphone HMMs with a single Gaussian density per state.

2.3. Lexica

Two different lexica were used to provide phonetic transcriptions of words:

- *USLex*: Pronunciations in the lexicon are based on a set of 45 phones. The lexicon was generated by merging different source lexica for American English (LIMSI '93, CMU dictionary, Pronlex). In addition, phonetic transcriptions for a number of missing words were generated by using the phonetic transcription module of the Festival speech synthesis system.

- *BEEPLex*: This lexicon was generated by exploiting the British English Example Pronunciations (BEEP) lexicon. Pronunciation models in this lexicon are based on a set of 44 phones. Transcription for a number of missing words were obtained by exploiting the pronunciation models in the *USLex* lexicon and mapping phonetic symbols into the BEEP phone set.

For each phone set and decoding pass, a set of state-tied, cross-word, gender-independent triphone HMMs were trained for recognition. Around 170,000 Gaussian densities, with diagonal covariance matrices, were allocated for each model set.

2.4. Language model

The language model was trained exploiting the language resources released for the IWSLT 2011 SLT Evaluation Campaign. These textual data consist of out of domain texts from several domains for a total of 763M words (after cleaning and removing double lines) plus a domain-specific corpus consisting of 2M words corresponding to subtitles of TED talks. Witten-Bell smoothing and mixture adaptation as supplied by the IRSTLM toolkit [6] were applied, using TED as adaptation data.

The LM is used twice: the first time to compile a static Finite State Network (FSN) which includes LM and lexicon and is used in the first and second decoding passes. The LM used for building this FSN is pruned in order to obtain an FSN of a manageable size, resulting in a recognition vocabulary of 200K words, 37M bigrams, 34M 3-grams, 38M 4-grams. The non-pruned LM is instead used to provide LM probabilities during word graph expansion.

2.5. Transcription process

In the IWSLT 2011 ASR evaluation, time boundaries of speech segments to be transcribed are given for each audio file. These non-overlapping speech segments are clustered by using a method based on the Bayesian information criterion [7]. The resulting clustering is exploited by the transcription system to perform cluster-based acoustic feature normalization and AM adaptation.

The first decoding pass is carried out with acoustic models based on *BEEPlex*, while the second and third decoding passes make use of acoustic models based on *USLex*. This configuration was chosen based on preliminary experiments on development data.

Cluster-based, text-independent acoustic feature normalization is first performed before HLDA projection. The output of the first decoding pass on these acoustic features is used as a supervision for conducting cluster-based CMLLR acoustic feature normalization and MLLR-based acoustic model adaptation [8] before the second decoding pass, where both the first-best output and word graphs are generated. The latter are expanded using the non-pruned four-gram language model described above and are compiled into corresponding decoding networks using the *USLex* lexicon. Also in this case, the best recognition hypothesis generated by the second decoding pass is exploited for conducting cluster-based CMLLR acoustic feature normalization and MLLR-based acoustic model adaptation before the third decoding pass.

2.6. System run

Transcription results, in terms of Word Error Rate (WER %), are given in Table 1. 18.8% and 18.2% WER were achieved by the primary FBK transcription system on the development (made by the union of the IWSLT 2010 development and evaluation sets) and evaluation sets, respectively. Results achieved by performing only the first two decoding passes are also reported.

Table 1: *WERs achieved by the primary FBK transcription system on the IWSLT 2011 ASR development and evaluation sets.*

System	Dev	Eval
Primary, step II	19.2	18.3
Primary, step III	18.8	18.2

3. MT and SLT Tasks

This year we continued work in the framework of phrase-based statistical machine translation. Our efforts mainly focused on domain adaptation by fill-up, a phrase table merging technique extensively described in [9] and presented in Section 3.2.2 of this paper.

We also explored the use of hybrid class-based language models to better capture the language style that is typical of public speeches. More details about this method and preliminary results are given in Section 3.2.4.

Concerning pre-processing, we addressed data redundancy in the monolingual corpora, experimented with a state-of-the-art Arabic-specific OOV handling tool, and constructed a custom punctuation insertion module for the post-processing of ASR outputs for translation.

3.1. General system descriptions

For both the Arabic-English and English-French systems, we set up a standard phrase-based system using the Moses toolkit [10]. The decoder features a statistical log-linear model including one or more phrase translation models, target language models, a phrase reordering model [11, 12], distortion, word and phrase penalties. In the Arabic-English task, we use a hierarchical reordering model [13], while in the English-French tasks we use a default word-based bidirectional model. The distortion limit is set to the default value of 6. As proposed by [14], statistically improbable phrase pairs are removed by all our phrase tables (before merging).

For each target language, two 5-gram language models are trained independently on the monolingual TED and NEWS datasets and are log-linearly combined at decoding time. The language models are trained with IRSTLM [6] with Modified Shift-Beta smoothing and no pruning. Additional experiments on class-based language models are performed in the Arabic-English task. The weights of the log-linear combination are optimized by means of a minimum error training procedure (MERT) [15].

The Arabic-English systems use cased translation models, while the English-French systems use lowercased models and a standard recasing post-process. The parallel training datasets used in our experiments are summarized in Table 2[2].

Table 2: *IWSLT11's TED task training corpora statistics.*

Corpus	Sentences	EN words	Avg sent.length
TED Ar-En	90K	1.7M	18.9
UN Ar-En	7.9M	220M	27.8
TED En-Fr	105K	1.9M	18.2
UN En-Fr	10.9M	251M	22.9
NEWS En-Fr	111K	2.6M	23.7

Concerning monolingual data, the usage of the English and French NEWS corpora for language modeling required specific pre-processing: due to heavy data redundancy and consequently abnormal n-gram counts, the application of Modified Shift-Beta smoothing on such data was impossible. We therefore removed all duplicate sentences from the corpora prior to language model estimation. Table 3 shows the percentage of redundant sentences within each NEWS sub-corpus, in both target languages. Surprisingly 72.83% of the English sentences and 52.55% of the French sentences were removed by this process, that is from 113M to 31M and from 25M to 12M sentences, respectively.

[2] The Europarl corpus was also available for English-to-French, but we did not use it in our experiments.

Table 3: *Percentage of redundant sentences from the English and French monolingual NEWS data.*

News source	EN % redundant	FR % redundant
News 2007	74.92	44.34
News 2008	71.35	51.28
News 2009	72.07	52.04
News 2010	74.72	51.39
News 2011	75.61	77.68
News Commentary v6	0.42	0.42
NEWS (all)	72.83	52.55

3.2. Arabic-to-English MT

3.2.1. Data processing

While English is pre-processed by a standard tokenizer, for Arabic we use our in-house tokenizer that also removes diacritics and normalizes special characters and digits. The Arabic text is then segmented with AMIRA [16] according to the Arabic Treebank (ATB) scheme that isolates conjunctions *w*+ and *f*+, prepositions *l*+, *k*+, *b*+, future marker *s*+, pronominal suffixes, but not the article *Al*+.

Considering the informal style of the TED talks, we designed a simple linguistic post-processing to apply a set of common contractions to the text. The rules mostly involve negations of auxiliary verbs, such as:

do not	→ *don't*
are not	→ *aren't*
have not	→ *haven't*
would not	→ *wouldn't*

3.2.2. Phrase table fill-up

The available parallel data come from two sources: a rather small in-domain corpus of TED talks and a large out-of-domain corpus of UN proceedings (see Table 2). The UN corpus has clearly a good potential in terms of model coverage. At the same time, it may be very noisy, given its domain and poor level of parallelism. We are also concerned with the large difference in size between the two corpora that may cause good TED translations to be drowned in the UN space.

We choose to combine the datasets by phrase table *fill-up*, a technique that exploits background knowledge to improve model coverage, while preserving the more reliable information of the in-domain corpus. The idea of fill-up goes back to Besling and Meier [17], which addressed the problem of LM adaptation for speech recognition, and was recently introduced in SMT by Nakov [18] and [9].

First, separate translation models are built from in-domain (TED) and background (UN) data. The background table is then merged with the in-domain table by adding only new phrase pairs that do not appear in the in-domain table. To keep track of a phrase pair's provenance, a binary feature is added that fires if the phrase pair comes from the background table. The resulting model can be tuned as usual, with the last feature acting as a scaling factor for out-of-domain translation scores.

In [9], fill-up is shown to perform similarly to or better than both off-line linear interpolation and decoding-time log-linear combination of multiple phrase tables. In comparison to log-linear combination, fill-up leads to faster convergence in MERT. At the same time, it doesn't require any additional tuning procedure, unlike linear interpolation.

Following [9]'s findings, we prune the background phrase pairs with more than four source words. We also merge our system's reordering models [11, 12, 13] using the same technique, with the only difference that no additional feature is introduced.

Table 4 shows the performance of fill-up versus other data combination techniques, namely concatenation of corpora, uniformly weighted linear interpolation and decoding-time log-linear combination of multiple phrase tables. Fill-up yields the best scores, followed by log-linear interpolation, for which MERT didn't converge in 25 iterations (the default maximum) probably due to the large number of weights to tune with two phrase tables and two reordering models. Linear interpolation with uniform weights obtained a noticeably lower score, and concatenation of corpora performed even worse than a system trained only on in-domain data.

Table 4: *%Impact of fill-up and other data combination techniques on Arabic-English MT performance (BLEU|NIST).*

Transl.model	test2010
only TED	24.96 | 6.434
concat(TED+UN)	23.45 | 6.130
linear(TED+UN)	25.15 | 6.401
logli(TED+UN)*	25.62 | 6.474
fillup(TED+UN)	**25.88** | **6.512**

*MERT didn't converge by the 25th iteration.

3.2.3. Handling Arabic OOV

Due to the rich morphology of Arabic, a high rate of out-of-vocabulary words (OOV) is generally observed in Arabic-to-English translation. To this end, we experimented with REMOOV [19], a tool specifically designed to handle Arabic OOVs in phrase-based SMT, through different techniques. The first technique, morphology expansion, tries to match the OOV with an in-vocabulary word that is a possible morphological variant of the OOV. The second, spelling expansion, tries to match the OOV with a known word that may be a possible correct spelling of the OOV. The third uses simple heuristics to detect numbers and non-Arabic words that should be output in the translation as is, while other OOVs are dropped (option *-drop-unknown* in Moses). We didn't use the other two techniques, namely dictionary expansion and name transliteration, because they employ built-in dictionaries that were not allowed for the evaluation. However, in experiments not reported here, we noted that these modules did not improve the performance of a TED-only system, but actually degraded it slightly.

The phrase pairs produced by REMOOV (through morphology and spelling expansion, and non-Arabic detection) are plugged into the system as an additional translation model. The feature weights are the same as those of the main translation model, except for the phrase penalty which is set to a very low value (-5).

The impact of REMOOV on our system's performance is not noticeable in terms of BLEU or NIST. A minor gain of +0.05% BLEU was observed only the devset, while on both test2010 and test2011 the use of REMOOV resulted in slightly lower scores. Yet these translations appear more readable for a human at a manual inspection. We then decided to submit this as a contrastive run (see *C1* in Table 6).

3.2.4. Hybrid class-based LM

One of the TED task's challenges is to capture the informal style of its speeches. We propose a technique to better exploit the in-domain data at this end, without the need of additional linguistic resources.

We can assume that the informal spoken register at issue is well-represented in the in-domain corpus, but not in the other corpora available for target language modeling. At the same time, we know that the TED corpus alone cannot ensure sufficient coverage to our system, given the variety of topics covered by the talks. The addition of background data can certainly improve the coverage and thus the fluency of our translations, but it may also move our system towards an insuitable register, such as that of written news.

In addition to conventional word-based n-gram LMs, we propose to use a hybrid class-based LM where only topic-specific words are mapped to classes. With an approximation, we use token frequency to detect such words, but other measures like document frequency may be used as well. Words that have frequency lower than the chosen threshold F in the in-domain corpus are replaced deterministically by their most likely Part-of-Speech (POS) tag[3], while other words are left in the text as is. In our experiments, we set the threshold to 500 occurrences, so that 25% of the tokens – corresponding to 99% of the types – are mapped to POS classes (English TED corpus statistics). The data thus obtained is used to train a 7-gram model with Witten-Bell smoothing.

Here are some examples of English sentences used to train the LM, before and after processing:

This new economy is pretty indifferent to size and strength, which is what's helped men along all these years.
This new **NN** is pretty **JJ** to **NN** and **NN** , which is what's **VBD NNS IN** all these years.

And you'd think , "Can four simple laws give rise to that kind of complexity?"
And you'd think , " **MD** four simple **NNS** give **NN** to that kind of **NN** ?"

Now you laugh, but that quote has kind of a sting to it, right.
Now you **VB** , but that **NN** has kind of a **NN** to it, right.

Because two vocabularies with different distributions are mixed in this type of data, our model may tend to excessively encourage, in the same context, the production of POS-mapped words at the expense of unmapped frequent words. To counteract this effect, we use an adaptation technique called Minimum Discriminative Information (MDI) [21]. MDI adaptation consists in scaling the n-gram probabilities of a background LM so that they match the target unigram distribution. For our purpose, instead, we use a unigram distribution estimated on corrected counts (c^*) where each POS class is assigned the count of its most frequent word:

$$\forall g \in W_{\geq F} \cup P :$$

$$c^*(g) = \begin{cases} c(g) & \text{if } g \in W_{\geq F} \\ c(\hat{w_g}) & \text{if } g \in P \end{cases}$$

where g is an entry of the processed data vocabulary, composed of frequent words ($W_{\geq F}$) and POS-classes (P), and $\hat{w_g}$ is the most frequent word that was mapped to g.

The hybrid class-based LM is trained only on the TED corpus and is added to the log-linear decoder as an additional target LM[4]. Table 5 shows the impact of hybrid class-based LM (*hyb500*) estimated with and without MDI adaptation. For comparison, we also experiment with a POS-class model (*allPOS*) where all the words are mapped to their most likely POS tag.

We can see that the addition of a hybrid LM adapted on corrected counts results in a promising improvement on the test2010, from 25.88 to 26.38 BLEU, whereas the POS-class LM yields a loss with respect to the baseline.

Table 5: *Impact of class-based language models on Arabic-English MT performance (%BLEU|NIST scores).*

Class-based LM	test2010
none	25.88 | 6.512
allPOS	25.56 | 6.493
hyb500	25.90 | 6.518
hyb500 mdi	**26.38** | **6.572**

3.2.5. Submitted runs

We present here the official results obtained by our systems at the competition. Our primary system (P) includes a phrase table built by the fill-up of TED with UN data, a hierarchical model obtained with the same method, two standard 5-gram LMs (TED and NEWS) and a hybrid class-based 7-gram LM trained on TED. The first contrastive run (*C1*) is produced by a similar system augmented with the REMOOV phrase table described in Section 3.2.3. The second contrastive run

[3] POS tags were obtained with Tree Tagger [20].

[4] The implementation of class-based LM provided by the IRSTLM and MOSES toolkits was used to this end.

(*C2*) is also obtained with a system similar to the primary, but without a hybrid LM.

All submitted runs were post-processed with contraction rules (see Section 3.2.1) and standard de-tokenization.

Table 6: *FBK system runs submitted for the Arabic-English MT track: official %BLEU|NIST results.*

Run	Notes	test2010	test2011
P		**26.37** | **6.573**	**24.31** | **6.1022**
C1	+remoov	26.25 | 6.554	24.17 | 6.0647
C2	-hybLM	25.94 | 6.521	24.23 | 6.0391

As shown by Table 6, the primary system performs best on the official test2011, as it does on test2010. Despite the apparent improvement observed at our manual inspections, the use of REMOOV has a slightly negative effect on BLEU: -0.12% on test2010 and -0.14% on test2011. Concerning hybrid LM, we note that its BLEU impact is minor on test2011: +0.08% whereas on test2010 it is +0.43%, The NIST score appears to be better affected in this sense (+0.052 on test2010 and +0.063 on test2011). We are currently exploring ways to improve this language modeling technique.

3.3. English-to-French MT

Our English-French translation experiments closely follow the methodologies described in the Arabic-English task. For robustness purposes, we train lowercased models and a standard recaser to enable our MT systems to be used on ASR outputs.

3.3.1. Data processing

For the English-French task, we have several data sources available: TED transcripts, NEWS data, UN transcripts, and Europarl proceedings. After perplexity analyses, we decide to omit the use of Europarl data from our experiments. The statistics of the parallel training data are listed in the second half of Table 2.

Based on perplexity analyses we use only the TED and NEWS monolingual French data in our experiments. As shown in Table 3, approximately 53% of the sentences in the monolingual NEWS data are redundant. We prune the redundant sentences from the 2007-2011 data sets but maintain the original News Commentary data set due to its low redundancy.

Concerning preprocessing we apply standard tokenization to the English and French data.

3.3.2. Cascaded fill-up

In contrast to the Arabic-English task, more bilingual sources are available in the English-French task. After omitting the Europarl data, we are left with the in-domain TED data and two out-of-domain data sets: NEWS and UN. Similar to the Arabic-English section, we use fill-up techniques described in Section 3.2.2 to combine multiple phrase and reordering tables. The NEWS data is similar in size to the TED data and covers a small amount of additional vocabulary, while the UN data has a large vocabulary coverage but contains additional noise that is not as suitable for the TED translation task.

Since an order of utility can be established on the data sets, we first construct a fill-up model with the TED+NEWS data. We then augment the TED+NEWS fill-up model with the UN data by repeating the fill-up process. In both iterations of fill-up, we prune the background phrase pairs with more than four words, as described in [9]. The addition of NEWS and UN data introduces two new phrase table features with weights that scale the contribution of the phrase table entries from each data set. After tuning, we observe a hierarchy of preference among the data sources: in-domain TED data is preferred; NEWS data is subsequently preferred with a small penalty, while the UN data receives a larger penalty. Had we tried to perform the cascade in the reverse order by introducing the UN data first, we would have seen virtually no impact from the smaller NEWS data. We describe the results of our fill-up models in the following section.

3.3.3. Submitted Runs

We present here the results obtained by our systems during the competition. Our primary (*P*) and constrastive (*C*) results are reported in Table 7 and are compared to a simple baseline (*B*), consisting of TED-only phrase and reordering tables. Each system uses a log-linear combination of the TED and pruned NEWS data language models. Unknown words are dropped from the translations in the unofficial 2010 test set results.

Our primary system is the cascaded fill-up model consisting of the TED, NEWS, and UN data. Our contrastive runs are the TED+NEWS fill-up model and a log-linear combination of the two phrase and reordering tables. We observe that the TED+NEWS fill-up model (*C1*) performs comparably to its log-linear counterpart (*C2*) in terms of BLEU and NIST scores. We also observe a residual benefit of cascading the fill-up process with the UN data (*P*).

3.4. English-to-French SLT

3.4.1. Data description

In the English-French SLT task, audio recordings and reference transcripts are provided. Additionally, ASR 1-best and word lattices were provided by the organizers. Utilizing our lowercased MT systems and recaser, we use the provided 1-best automatic transcripts for model development. For our submission, we use the provided system combination of the systems labeled 0, 1, 2, and 4 from the ASR evaluation set.

Table 7: *FBK system runs submitted for the English-French MT track: official %BLEU and NIST evaluation results are reported for the 2011 test set. Primary (P) and contrastive (C) results are compared to an unsubmitted baseline (B).*

Run	Translation model	test2010	test2011
P	cascade-fill(TED+NEWS+UN)	**30.64** | **7.221**	**34.87** | **7.4169**
C1	fillup(TED+NEWS)	30.22 | 7.177	34.14 | 7.3231
C2	logli(TED+NEWS)*	30.29 | 7.154	33.93 | 7.2561
B	only TED	29.96 | 7.157	–

*MERT didn't converge by the 25th iteration.

3.4.2. Introducing punctuation

The reintroduction of punctuation is particularly important to our recasing model, as the recaser relies on punctuation to accurately recase words at the beginning of sentences. To introduce punctuation in ASR transcripts we exploited the *hidden-ngram* tool from the SRILM Toolkit[5], in combination with a hybrid class-based LM estimated on the TED data as described before.

Table 8 outlines the effects of our class-based punctuator against a naïve baseline that adds periods to the end of every line. Each system is trained on a baseline system with a TED-only translation model and a log-linear combination of TED and NEWS data as language models. Simply adding periods to the end of every line yields approximately a 9.9% BLEU improvement over a non-punctuated test2010 set. The class-based punctuator yields approximately a 12.5% relative BLEU improvement over the naïve punctuator for an overall 23.7% improvement versus no punctuation.

Table 8: *Effects of the class-based punctuator on the translation quality of our baseline SMT system (%BLEU|NIST). All systems use a TED-only translation model and a log-linear combination of TED and NEWS data as language models.*

Punctuation	test2010
No punctuation	17.70 | 5.451
Add periods to the end of every line	19.45 | 5.785
Class-based	21.89 | 6.055

3.4.3. Submitted Runs

We present here the results obtained by our systems during the competition. We use the same primary and contrastive systems in the SLT task as in the English-French MT task: a cascaded fill-up model, a TED+NEWS fill-up model, and a log-linear TED+NEWS model. We use the class-based punctuator described above on the lowercased Rover-constructed ASR system combination transcriptions from the test2011 data set. After generating lowercased translations, we pass the results through a standard recaser. Table 9 describes the performance of each of our submissions on the test2010 and test2011 data sets.

Table 9: *FBK system runs submitted for the English-French SLT track: official %BLEU and NIST evaluation results are reported for the 2011 test set.*

Run	Translation model	test2010	test2011
P	cascade-fill(TED+NEWS+UN)	**22.36** | **6.097**	**24.31** | **6.1457**
C1	fillup(TED+NEWS)	22.22 | 6.087	23.59 | 6.0309
C2	logli(TED+NEWS)	22.19 | 6.086	23.96 | 6.0579

We note that the lowercased, unpunctuated ASR outputs used in the SLT task yield BLEU scores that are about 8 and 10 points lower than the MT task for the respective 2010 and 2011 test sets (see Table 7). We also note a smaller difference in scores between our cascade fill-up and standard fill-up models.

4. Conclusions

We presented our submission runs to the IWSLT 2011 Evaluation Campaign for the ASR English track, the MT Arabic-English and English-French tracks, and the SLT English-French track.

Our ASR system was trained on a significant portion of TED talk recordings, by exploiting an automatic data selection method evaluating the fidelity of the provided transcripts. For the MT and SLT talks, we introduced phrase table fill-up models for domain adaptation and a hybrid class-based language model to adapt to the speaking style of talks.

An issue worth exploring in the future is topic adaptation, both on the ASR and MT sides. In the case of both ASR and MT, topic-based language model adaptation is suitable for this evaluation campaign. Additionally, phrase table topic adaptation should be plausible.

5. Acknowledgements

This work was partially supported by the T4ME network of excellence (IST-249119) and the EuroMatrixPlus project (IST-231720), both of which are funded by the European Commission under the Seventh Framework Programme for Research and Technological Development.

6. References

[1] M. Federico, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2011 Evaluation Campaign,"

[5] http://www.speech.sri.com/projects/srilm

in *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[2] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[3] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive Training Using Simple Target Models," in *Proc. of ICASSP*, Philadelphia, PA, March 2005, pp. I–997–1000.

[4] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization." *Computer Speech and Language*, vol. 20, no. 1, pp. 107–123, Jan. 2006.

[5] D. Giuliani and F. Brugnara, "Experiments on Cross-System Acoustic Model Adapatation," in *Proc. of ASRU Workshop 2007*, Kyoto, Japan, Dec. 2007, pp. 117–122.

[6] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proceedings of Interspeech*, Melbourne, Australia, 2008, pp. 1618–1621.

[7] M. Cettolo, "Segmentation, classification and clustering of an italian broadcast news corpus," in *Proc. of Content-Based Multimedia Inf. Access Conf. (RIAO)*, Paris, France, 2000, pp. 372–381.

[8] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[9] A. Bisazza, N. Ruiz, and M. Federico, "Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation," in *SUBMITTED TO International Workshop on Spoken Language Translation (IWSLT)*, 2011.

[10] 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 Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180. [Online]. Available: http://aclweb.org/anthology-new/P/P07/P07-2045.pdf

[11] C. Tillmann, "A unigram orientation model for statistical machine translation," in *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2004.

[12] 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 the International Workshop on Spoken Language Translation*, October 2005.

[13] M. Galley and C. D. Manning, "A simple and effective hierarchical phrase reordering model," in *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2008, pp. 848–856.

[14] H. Johnson, J. Martin, G. Foster, and R. Kuhn, "Improving translation quality by discarding most of the phrasetable," in *In Proceedings of EMNLP-CoNLL 07*, 2007, pp. 967–975.

[15] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, E. Hinrichs and D. Roth, Eds., 2003, pp. 160–167. [Online]. Available: http://www.aclweb.org/anthology/P03-1021.pdf

[16] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," in *HLT-NAACL 2004: Short Papers*, D. M. Susan Dumais and S. Roukos, Eds. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 149–152.

[17] S. Besling and H. Meier, "Language model speaker adaptation," in *Proceedings of the 4th European Conference on Speech Communication and Technology*, vol. 3, Madrid, Spain, 1995, pp. 1755–1758.

[18] P. Nakov, "Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing. ," in *Workshop on Statistical Machine Translation, Association for Computational Linguistics*, 2008.

[19] N. Habash, "Remoov: A tool for online handling of out-of-vocabulary words in machine translation," in *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, K. Choukri and B. Maegaard, Eds. Cairo, Egypt: The MEDAR Consortium, April 2009, pp. 217–220.

[20] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[21] M. Federico, "Efficient language model adaptation through MDI estimation," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 1583–1586.