FBK's Machine Translation Systems for IWSLT 2012's TED Lectures

N. Ruiz, A. Bisazza, R. Cattoni, M. Federico

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

Abstract

This paper reports on FBK's Machine Translation (MT) submissions at the IWSLT 2012 Evaluation on the TED talk translation tasks. We participated in the English-French and the Arabic-, Dutch-, German-, and Turkish-English translation tasks. Several improvements are reported over our last year baselines. In addition to using fill-up combinations of phrase-tables for domain adaptation, we explore the use of corpora filtering based on cross-entropy to produce concise and accurate translation and language models. We describe challenges encountered in under-resourced languages (Turkish) and language-specific preprocessing needs.

1. Introduction

FBK's machine translation activities in the IWSLT 2012 Evaluation Campaign [1] focused on the speech recognition and translation of TED Talks[1], a collection of public speeches on a variety of topics and with transcriptions available in multiple languages. In this paper, we discuss our involvement in the official Arabic-English and English-French Machine Translation tasks, as well as the auxillary German-English, Dutch-English, and Turkish-English Machine Translation tasks.

We begin with an overview of the research procedure in common with all of language pair experiments in Section 2: namely, data filtering, phrase and reordering table fill-up, and mixture language modeling. In Section 4 we discuss our Arabic-English and Turkish-English MT systems. In Section 3 we discuss our English-French submissions. In Section 6 we discuss our German- and Dutch-English systems. Finally, in Section 8 we summarize our findings.

2. TED Machine Translation Overview

For all systems except for our Turkish-English system, we set up a standard phrase-based system using the Moses toolkit [2]. We construct a statistical log-linear model including a filled-up phrase translation and hierarchical reordering models [3, 4, 5], a primary mixture target language model (LM), as well as distortion, word, and phrase penalties. The distortion limit is set to the default value of 6, except for Arabic- and Turkish-English (see respective sections). As proposed by [6], statistically improbable phrase pairs are removed from our phrase tables.

For each target language, we train 5-gram mixture language models from the available corpora, as described in Section 2.3. The language models are trained with IRSTLM [7] with improved Kneser-Ney smoothing and no pruning. Additional experiments on hybrid word/class language models are performed in the Arabic-English task. The weights of the log-linear combination are optimized via minimum error rate training (MERT) [8].

In the following sections, we discuss the data selection, phrase and reordering table fill-up, and mixture language modeling used by each of our systems. We follow the discussion with our language-specific submissions.

2.1. Data selection

Each out-of-domain corpus was domain-adapted by filtering aggressively using a cross-entropy difference scoring technique described by [9] on the target side and optimizing the perplexity against the (target language) TED training data by incrementally adding sentences.

The idea of data selection is to find the subset of sentences within an out-of-domain corpus that better fits with a given in-domain corpus. Each sentence of the out-of-domain corpus is evaluated by comparing its likelihood (in terms of cross-entropy) to appear in the out-of-domain corpus against its likelihood to compare in the in-domain corpus. In order to decide how many sentences to keep, we build an out-of-domain language model incrementally and measure its perplexity on the in-domain TED data. The two language models we compare are built from the same dictionary, namely the in-domain words occurring more than a specified frequency. All other words in the in-domain and out-of-domain corpora are taken as out-of-vocabulary words. For this kind of problem it is generally sufficient to work with 3-grams language models estimated on words occurring at least twice in the in-domain set.

Figure 1 shows the effects of data selection on the four out-of-domain corpora used for language modeling in all of our foreign-to-English MT submissions. Three of the corpora are subcorpora drawn from seven available news text sources in the LDC English Gigaword (Fifth Edition) corpus.

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

The statistics of each corpus are shown in Table 1.

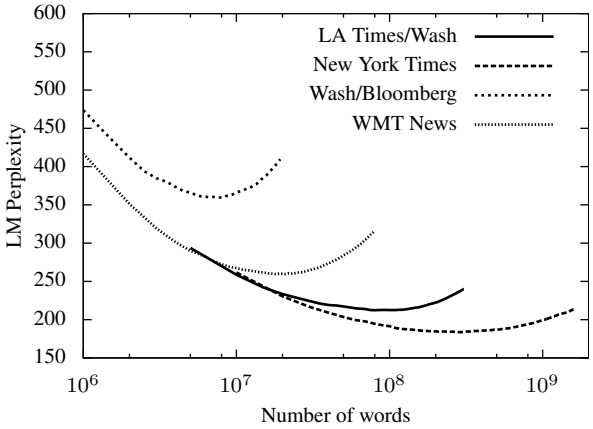

Figure 1: Effects of cross-entropy data selection on perplexity (PP) for the English monolingual out-of-domain data used by all foreign-to-English systems. Sentences are incrementally added based on their rank with trigram PP measures reported against the IWSLT 2010 TED development set. The PP scores reach a saddle point in which the inclusion of additional sentences worsens the language model. Each LM requires only a fraction of the entire available corpus.

	Unfiltered		Filtered		
Corpus	**Lines**	**Tokens**	***Lines**	***Tokens**	**% Filt**
Gigaword LAT	6.73M	312M	1.6M	80M	74.4
Gigaword NYT	38.7M	1.6B	6.75M	300M	81.3
Gigaword WP	421K	19.8M	135K	7M	64.6
WMT News	31M	849M	878K	20M	97.6

Table 1: Filtering statistics on the monolingual English (sub)corpora used in FBK's systems. Sentences were incrementally added until a local minimum perplexity value against the development set was reached.

2.2. Phrase table fill-up

As we did last year, we combine phrase tables via fill-up [10, 11]. Using the recommendations of [11], we add k-1 binary provenance features for each of the k phrase tables to combine. Treating the TED phrase table as in-domain, we merge out-of-domain phrase pairs that do not appear in the in-domain TED table, along with their scores. Moreover, out-of-domain phrase pairs with more than four source tokens are pruned. The fill-up process is performed in a cascaded order, first filling in missing phrases from the corpora that are closest in domain to TED.

2.3. Mixture language model adaptation

After performing data selection and cross-entropy filtering on the provided monolingual corpora, we perform LM domain adaptation via mixture modeling [12].

For our foreign-to-English MT submissions, we construct a common 5-gram mixture LM consisting of TED data, a subset of corpora from the LDC Gigaword fifth edition corpus, and the WMT News Commentary. From the Gigaword corpus, we select the articles from the Los Angeles Times/Washington Post, New York Times, and Washington Post/Bloomberg subcorpora. After performing cross-entropy filtering on each subcorpus, we perform mixture model adaptation with the TED corpus as the in-domain background. French language model statistics are reported in Section 3.3.

3. English-French

More monolingual and parallel data were available in the English-French translation task. Several of the corpora were too large and noisy to use efficiency, which underscored the necessity of data selection and filtering. In the following sections we discuss the data selection, phrase and reordering table fill-up, and mixture language modeling approaches used for our English-French MT systems and report results on the official test sets.

3.1. Data selection

We perform data selection using the cross-entropy filtering technique described above, both for language and for translation modeling. In order to filter parallel corpora, we apply the cross-entropy filtering technique on the French (target-side) texts and prune the corresponding English segments. Table 2 provides statistics on the preprocessed monolingual and parallel corpora used by our systems, before and after filtering. In both monolingual and parallel corpora we observe over a 85% reduction in the number of words by filtering.

	Unfiltered		Filtered		
Corpus	**Lines**	**Tokens**	***Lines**	***Tokens**	**% Filt**
Europarl	2.0M	61.9M	200K	4.2M	93.2
Giga French	19.7M	570M	1.08M	25.5M	96.6
Gigaword AFP	18.3M	668M	1.08M	46.1M	93.1
Gigaword APW	6.5M	255M	660K	34.7M	86.4
MultiUN	10.5M	290M	228K	5.2M	98.2
WMT News	7.5M	182M	900K	20.9M	88.5

Table 2: French filtering statistics on the tokenized and cleaned (sub)corpora used in FBK's systems. Europarl, Giga French, and MultiUN were used for translation model training, while French side of the Giga corpus and the monolingual Gigaword AFP and WMT News corpora were used for language model training.

3.2. Phrase table

More parallel data was available in the English-French translation task than the other MT tracks. In particular, the MultiUN and Giga French corpora were too large and noisy to use reliably for translation modeling without filtering. Table 2 shows that the size of these corpora were reduced by over 95% using cross-entropy filtering.

We use the filtered TED, Europarl, MultiUN, and Giga French parallel corpora for translation model training. Our experiments from last year showed little improvement from

using the parallel WMT News Commentary corpus. In order to reduce the size of the translation models and to stabilize MERT behavior, we independently train phrase and reordering tables on each corpus and experiment with several fill-up configurations with the TED as the in-domain corpus. Table 3 lists BLEU and TER evaluation results[2] on the IWSLT 2010 TED test set, three independent MERT runs for each fill-up combination. Each system uses the mixture LM described later in Section 3.3. In particular, we do not see any significant improvements filling up with using Europarl or MultiUN, but rather with the Giga French corpus. In order to improve the coverage of the TED and Giga fill-up models, we cascaded fill-up with Europarl and MultiUN respectively. While we do not observe significant improvement with the cascaded fill-up from Table 3, we later observe different results on our submitted runs.

	BLEU ↑			TER ↓		
System	**Avg**	$\overline{s}_{\textbf{sel}}$	p	**Avg**	$\overline{s}_{\textbf{sel}}$	p
TED-only	32.2	0.5	-	49.7	0.5	-
Fill(TED+Euro)	32.3	0.5	0.27	49.5	0.5	0.03
Fill(TED+UN)	32.2	0.5	0.60	49.4	0.5	0.00
Fill(TED+Giga)	32.5	0.5	0.03	49.4	0.5	0.01
Fill(TED+Giga+UN)	32.4	0.5	0.09	49.6	0.5	0.14
Fill(TED+Giga+Euro)	32.4	0.5	0.12	49.5	0.5	0.03

Table 3: Evaluation of phrase table combinations on the IWSLT 2010 TED test set, averaged across three MERT runs. Each translation system uses the mixture LM described in Section 3.3. Phrase tables are filled-up in a left-to-right order. p-values are relative to the system trained with only the TED phrase table. $\overline{s}_{sel}$ indicates the variance due to test set selection.

3.3. Language modeling

In order to determine which monolingual data to use for language modeling, we trained 5-gram language models on each unfiltered corpus and evaluated their perplexity scores on the in-domain TED development data. From our experiments last year, the monolingual WMT News Commentary corpus yielded well-performing LMs. The Gigaword corpus consisted of articles from the Agence France-Presse (AFP) and Associated Press Worldstream (APW) newswires. Our perplexity analyses showed that APW did not model the TED domain well; thus, we opt to omit it. To our surprise, the French side of the parallel Giga French corpus modeled the TED domain well after filtering – even better than the TED training data!

Rather than log-linearly combining four distinct LMs and optimizing four feature weights, we combine the LMs with mixture modeling and evaluate their cumulative effects on the IWSLT 2010 development set in Table 4. After confirming that the four LMs in combination improve perplexity, we construct a 5-gram mixture model. Table 5 suggests that the mixture LM alone is responsible for a 2.7 BLEU improvement over a TED-only 5-gram baseline.

[2]Evaluation results were performed with MultEval v0.3 [13].

Corpora	**PP dev$_{2010}$**	**% OOV**
TED	139.40	1.65%
Giga-EF	126.65	0.85%
TED + Giga-EF	85.60	0.7%
+ Gigaword AFP	81.34	0.4%
+ WMT News	80.19	0.4%

Table 4: Perplexity of 3-gram mixture LMs evaluated on the IWSLT 2010 development set. Giga French, Gigaword AFP, and WMT News corpora are incrementally added to the in-domain TED training corpus and provide excellent coverage of the development data.

PT	**LM**	**Metric**	**Opt 1**	**Opt 2**	**Opt 3**	**Avg**
TED	TED	BLEU	29.75	29.95	29.72	29.74
		NIST	7.167	7.184	7.178	7.170
	Mix	BLEU	32.37	32.44	32.44	32.42
		NIST	7.463	7.438	7.438	7.443

Table 5: Effects of mixture LM on the IWSLT 2010 TED test set. Results are calculated across three MERT optimizations with their weights averaged for final evaluation. The mixture LM results in roughly 2.7 BLEU and 0.27 NIST improvements against a TED-only phrase table.

3.4. Submitted runs

Our primary (*P*) and constrastive (*C*) results are reported in Table 6 and are compared to a simple TED baseline (*B*), consisting of TED-only phrase and reordering tables. All systems use the mixture LM described in the previous section. Each system's feature weights are averaged over three MERT optimizations. The fill-up model with Europarl yielded higher BLEU and NIST scores on both the 2010 development and test sets; thus by providing additional phrase coverage we opted to submit it as our primary system. Our TED+Giga fill-up system served as our contrastive baseline. Each system performed similarly on the official test sets, though the MultiUN filled-up model was not consistent across the different test sets. Our primary system performed equally with our contrastive baseline on the 2011 test set in terms of BLEU, but performed slightly (though not significantly) worse in terms of NIST, while on the 2012 test set we observe a 0.3 BLEU improvement.

	PT	**Metric**	**dev$_{2010}$**	**tst$_{2010}$**	**tst$_{2011}$**	**tst$_{2012}$**
B	TED	BLEU	27.71	32.22	–	–
		NIST	6.600	7.397	–	–
P	Fill(TED	BLEU	28.42	32.42	37.43	37.29
	+Giga+Euro)	NIST	6.697	7.443	7.713	8.039
C_1	Fill(TED	BLEU	28.11	32.39	37.43	36.99
	+Giga)	NIST	6.660	7.450	7.737	8.024
C_2	Fill(TED	BLEU	28.23	32.52	37.36	37.24
	+Giga+UN)	NIST	6.681	7.460	7.715	8.051

Table 6: Results of submitted runs evaluated on the IWSLT TED development and test sets. Evaluation on the 2010 data sets are compared against a TED-only phrase table. All systems use the mixture LM described in Section 3.3. MT system weights are averaged across three MERT optimizations for final evaluation.

4. Arabic-English

The Arabic-English language pair is characterized by notable differences in morphological richness and word order. We follow last year's experience to deal with morphology and address word reordering by using an improved version of the distortion penalty that was proposed by [14]. In addition to that, we integrate a hybrid class language model [15] that proved to improve our system of last year.

4.1. Preprocessing

For Arabic we use our in-house tokenizer that also removes diacritics and normalizes special characters and digits. Then, segmentation is performed by the AMIRA toolkit [16] based on SVM classifiers, 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+*. Arabic training data statistics are given in Table 7.

		AR tokens		EN tokens
Corpus	**Lines**	**unsegm.**	**Amira-segm.**	
TED	137K	2.1M	2.5M	2.7M
MultiUN	8M	188M	224M	220M

Table 7: Arabic-English training data statistics showing number of Arabic tokens before and after segmentation.

4.2. Phrase table

While word alignment is obtained on the union of all available data, the translation model is built by filling up a TED-only phrase table with a MultiUN-only phrase table. As previously said, out-of-domain (MultiUN) phrase pairs with more than four source words are filtered out. The lexicalized reordering table is obtained with the same procedure.

4.3. Early distortion cost

Moore and Quirk [14] proposed an improvement to the distortion penalty used in Moses, which consists in "incorporating an estimate of the distortion penalty yet to be incurred into the estimated score for the portion of the source sentence remaining to be translated." The new distortion penalty has the same value as the usual one over a complete translation hypothesis (provided that the jump from the last translated word to the end of the sentence is taken into account). As a difference, though, it anticipates the gradual accumulation of the total distortion cost making partial translation hypotheses with the same number of covered words more comparable with one another. We have implemented this 'early distortion cost' option in the Moses platform and used it in our systems. As shown in Table 8, increasing the distortion limit from the default value of 6 to 8 has normally a negative impact because standard distortion does not properly control long jumps. On the contrary, when *early* distortion cost is used, a slightly higher distortion limit is preferable, yielding an improvement of +0.2 BLEU and +0.04 NIST over the baseline.

DL	**DC**	**BLEU**	**NIST**
6	std	26.12	6.514
8	std	25.95	6.460
8	**edc**	26.31	6.551

Table 8: Effects of distortion limit (DL) and distortion cost (DC), standard or early, on the IWSLT 2010 TED test set.

4.4. Mixture language modeling

In Arabic-English too, we use mixture modeling for domain adaptation. Concerning data selection, we find that a 4-gram LM trained on unfiltered data performs slightly better in terms of BLEU than the filtered 5-gram LM presented in section 2.3 (see first two rows of Table 9). A possible explanation is that, if translation gets more difficult, especially due to reordering, relying on a much larger number of n-grams helps to discriminate correct versus incorrect phrase concatenations. This discrimination capability may not reflect on the perplexity, which only measures how a LM predicts correct text. Thus, we use the unfiltered LM for the Arabic-English systems. It should be noted, though, that this model requires twice as much memory to function.

LM	**BLEU**	**NIST**
MixFiltered.5g	25.92	6.465
MixAll.4g	26.31	6.551
MixAll.4g + TED.Hybrid10g	26.65	6.591

Table 9: Effects of data selection and hybrid language modeling on the IWSLT 2010 TED test set.

4.5. Hybrid language modeling

In addition to the mixture model, we use an in-domain hybrid word/class LM that was proposed by [15] to address style adaptation when out-of-domain data is likely to bias the system towards an unsuitable language style (e.g. news versus talks). Following the paper, we train a high order (10-gram) LM on TED data where infrequent words were mapped to their most likely Part-of-Speech tags, and frequent words to their lemma. We set the frequency threshold so that 25% of the tokens – corresponding to about 2% of the types – are replaced by part-of-speech (POS) tags. Adding this model to the log-linear combination yields a gain of +0.3 BLEU and +0.04 NIST (see Table 9).

4.6. Submitted runs

Table 10 presents results of our baseline (B), primary (P) and contrastive (C) systems on the IWSLT 2010, 2011 and 2012 TED test sets. All Arabic-English systems use the same phrase and reordering models, obtained by fill-up of TED

and UN data. Our best submission is obtained with early distortion cost, a distortion limit of 8 words and an in-domain hybrid LM in addition to a large unfiltered mixture LM.

	LM	DL	Metric	tst_{2010}	tst_{2011}	tst_{2012}
B	MixAll.4g	6	BLEU	26.12	–	–
			NIST	6.514	–	–
P	MixAll.4g	8	BLEU	26.65	25.46	27.86
	+TED.Hybrid10g	[edc]	NIST	6.591	6.232	6.881
C_1	MixAll.4g	8	BLEU	26.31	25.19	27.74
		[edc]	NIST	6.551	6.205	6.903
C_2	MixFiltered.5g	8	BLEU	26.11	25.13	27.54
	+TED.Hybrid10g	[edc]	NIST	6.520	6.190	6.828

Table 10: Results of Arabic-English submitted runs evaluated on the IWSLT TED development and test sets.

5. Turkish-English

The additional training data provided for this language pair was limited to the South European Times news corpus. In our experiments we found that this data was not helpful for translation modeling and decided to use it only for word alignment[3]. A reason for this could be the size of this corpus – only slightly larger than the TED data – that is enough to bring noise into the system but not enough to improve its coverage in a significant way.

We then focus on preprocessing techniques to address the agglutinative Turkish morphology and evaluate the performance of phrase-based against hierarchical systems.

5.1. Morphological segmentation

Turkish preprocessing involves supervised morphological analysis [17] and disambiguation [18], followed by selective morpheme segmentation as described in [19]. We compare two of the segmentation schemes that were proposed and tested on the BTEC task by [19] and [20]:

- **'MS6'** deals only with nominal suffixes (case and possessive),
- **'MS15'** deals with nominal suffixes and verbal suffixes (copula, person subject, negation, ability, passive and causative suffixes).

The latter segmentation scheme is more aggressive, which is good for model coverage but can make the translation harder (especially the reordering problem, due to the larger number of possible input permutations).

To evaluate the actual importance of supervised methods, we also build a contrastive system using a fully data-driven segmentation approach proposed by [21] and implemented in the Morfessor Categories-MAP software. We train Morfessor on the TED training corpus, and obtain a unique segmentation of each word type into a sequence of morpheme-like units (*morphs*). As an intermediate solution between words and morphs – which are typically rather short – we concatenate the sequence of non-initial morphs to form so-called *word endings*[4]. In this way, each word can be segmented into at most two parts.

		TR tokens				EN
Corpus	Lines	unsegm.	MS6	MS15	Morf.	tokens
TED	125K	1.8M	2.0M	2.2M	2.4M	2.4M

Table 11: Turkish-English training data statistics showing how the number of Turkish tokens varies according to the segmentation method: supervised (MS6 and MS15) or unsupervised (Morfessor).

Turkish training data statistics in different segmentation settings are given in Table 11, while the effect on translation quality is shown in Table 12. Notice the very high distortion limit chosen because of the important order differences between English and Turkish, a head-final SOV language. In this set of experiments we use a 4-gram mixture LM trained on unfiltered data. The results show that supervised segmentation (MS15) can noticeably outperform the unsupervised one (Morfessor *word endings*), but they also show that the choice of a particular segmentation scheme is very important. In fact, the supervised MS6 scheme does no better than the unsupervised. We decide to use MS15 for the rest of the evaluation, however it is possible that the unsupervised approach may be improved by devising other ways to recombine the morphs.

DL	DC	Segment.	BLEU	NIST
15	std	MS6	13.61	5.280
15	std	MS15	14.38	5.273
15	std	unsup.	13.45	5.080
15	**edc**	MS15	14.53	5.299

Table 12: Effects on translation quality (IWSLT 2010 test set) of Turkish morphological segmentation, and of standard versus *early* distortion cost (see Section 4.3).

5.2. Translation model: phrase-based vs. hierarchical

As we only use TED training data, no adaptation technique is required for translation modeling.

Given the global and hierarchical nature of word reordering patterns in this language pair, we thought that a hierarchical translation system [23] could work better than a regular phrase-based one. We then construct a rule table with maximum rules span 15 and Good Turing score smoothing, and switch to chart decoding (all within the Moses platform).

The hierarchical system strongly outperforms the phrase-based one, with a +1.7 BLEU and 0.25 NIST gain (see Table 13) proving the complexity of the word reordering problem in Turkish-English.

[3] We concatenated the two corpora, ran GIZA++ on them, but only used the TED portion of the result.

[4] This approach is sometimes adopted in language modeling for Turkish speech recognition, see for instance [22].

5.3. Submitted runs

We submitted two systems: the hierarchical as primary and the phrase-based with early distortion cost and a high distortion limit (15) as contrastive. Both of our official systems include a 6-gram mixture LM trained on the filtered data described in Section 2.1.

	System	Segm.	Metric	tst_{2010}	tst_{2011}	tst_{2012}
P	hierarchical	MS15	BLEU	16.61	17.24	17.15
			NIST	5.570	5.560	5.702
C	phrase-based (dl=15, edc)	MS15	BLEU	14.92	15.45	15.24
			NIST	5.318	5.289	5.145

Table 13: Results of Turkish-English submitted runs evaluated on the IWSLT TED development and test sets.

6. German-English

Translating German compound words (also known as "compounds") is a challenge for Machine Translation: the first subsection focuses on the experiments we performed on compounds splitting. We subsequently report on the translation and language models used in our submissions and present our system results on the official test sets.

6.1. Word splitting

In order to choose the best splitter sub-system, we performed some preliminary experiments. We use the splitting tool provided in Moses (see [24]), which is based on a trainable model. We test several splitter configurations with models trained on all the German data available for the MT track of the TED Task, but with different filtering techniques and parameter settings, inspired by [25]). For the sake of efficiency, we perform the experiments on the TED corpora (namely the provided training and 2010 development and test sets). After applying a standard tokenization step, different groups of data sets are obtained, one for each splitting configuration.

We conduct two sets of experiments; in the first we compute the perplexity and OOV-rate on the dev and test sets using the LM learned on the training set, while in the second we build SMT systems for each splitting configuration and evaluate their translations. It is worth noting that the splitters work only on the source language and do not affect the target language (English).

Table 14 lists the outcomes of the first set of experiments: the *normal* splitter utilizes the default parameter setting of the tool, while in the *aggressive* splitter we change the parameters to allow decomposition into short words (minimum 2 characters). The best performance in terms of perplexity and OOV-rate reduction is exhibited by the aggressive splitter.

There are no statistically significant differences among the translations provided by the three systems (unsplitted, normal- and aggressive-splitting). This can be explained mainly by the limited size of the training set. In the same experiments performed with all the available German data, we observe a marginal but statistically significant improvement on translation scores when performing both normal and aggressive splitting.

Split	Set	Tokens	Voc	Perplexity	OOV%
no	training	2419470	101623	–	–
	dev_{2010}	19082	4194	556.26	3.15
	tst_{2010}	30316	5181	417.11	2.66
normal	training	2474654	78113	–	–
	dev_{2010}	19444	4160	497.21	2.37
	tst_{2010}	30924	5072	377.40	1.85
aggressive	training	2508243	72091	–	–
	dev_{2010}	19725	4140	464.94	2.11
	tst_{2010}	31312	5027	355.26	1.62

Table 14: Statistics on the German TED sets obtained by varying the splitting configuration. The *aggressive* splitter exhibits the best performance in terms of perplexity and OOV-rate reduction.

6.2. Phrase table

For translation modeling we use the four provided data sets. The MultiUN bilingual entries are obtained by aligning parallel documents at sentence level with the Hunalign 1.1 tool [26] after standard tokenization. The statistics of the tokenized unsplitted corpora are shown in Table 15.

Corpus	Lines	DE tokens	EN tokens
TED	130K	2.4M	2.6M
news-commentary-v7	159K	4.0M	3.9M
MultiUN	163K	5.6M	5.6M
europarl-v7	1.9M	50.5M	53.0M

Table 15: German-English parallel training corpora statistics.

While word alignment is obtained on the union of all available data, the translation model is built by filling up a TED-only phrase table with two other phrase tables: the former obtained from WMT News Commentary v7 corpus and the latter from the union of MultiUN and Europarl v7 corpora. This partition has been chosen to maximize domain homogeneity in the three sub-corpora. The lexicalized reordering table is obtained with the same procedure.

6.3. Submitted runs

Table 16 presents results of our primary (P) and contrastive (C) systems on the IWSLT 2010, 2011 and 2012 TED test sets. Both systems use the English 5-gram mixture LM previously described in section 2.3 and differ only on the word splitting technique. Evaluation scores are rather close; the aggressive splitter appears to exhibit slightly better (although not statistically significant) performance.

7. Dutch-English

In the following sections we present the systems developed for the Dutch-English MT track of the TED task.

	Splitter	Metric	tst_{2010}	tst_{2011}	tst_{2012}
P	aggressive	BLEU	29.36	32.38	28.17
		NIST	7.257	7.513	7.004
C	normal	BLEU	29.49	32.13	28.12
		NIST	7.224	7.447	7.003

Table 16: Results of submitted runs evaluated on the German-English IWSLT TED development and test sets.

7.1. Word splitting

Like German, the Dutch language includes compounds. However, no specific splitting experiments were performed on Dutch: as splitters, we ported into Dutch the best splitting configurations found in our German experiments. The splitting models were trained on all available Dutch corpora.

7.2. Phrase table

For translation modeling, we use both the TED and Europarl v7 corpora. The statistics of the tokenized unsplit corpora are shown in Table 17.

Corpus	Lines	NL tokens	EN tokens
TED	128K	2.3M	2.5M
europarl-v7	2.0M	55.3M	54.8M

Table 17: Dutch-English parallel training corpora statistics.

Word alignment is obtained on the concatenation of both corpora. The translation model is built by filling up the TED-only phrase table with the out-of-domain Europarl phrase table. The same procedure is applied for the lexicalized reordering table.

7.3. Submitted runs

Table 18 presents results of our primary (P) and contrastive (C_1 and C_2) systems on the IWSLT 2010, 2011 and 2012 TED test sets. The three systems differ in the splitters (normal for P and C_1, aggressive for C_2) and language models: all of them use the English mixture LM previously described in section 2.3, but differ in length (4-gram for P, 5-gram for C_1, 6-gram for C_2). The evaluation scores do not highlight a single outperforming system.

	Splitter	Metric	tst_{2010}	tst_{2011}	tst_{2012}
P	normal	BLEU	33.85	36.11	32.68
		NIST	7.763	7.921	7.743
C_1	normal	BLEU	33.91	36.23	32.48
		NIST	7.759	7.946	7.722
C_2	aggressive	BLEU	33.84	35.82	32.68
		NIST	7.726	7.881	7.725

Table 18: Results of submitted runs evaluated on the Dutch-English IWSLT TED development and test sets.

8. Conclusions

We presented our submission runs to the IWSLT 2012 Evaluation Campaign for the TED MT tracks. Our MT systems benefited most from data filtering techniques and mixture language modeling. In particular, we observed significant BLEU improvements using mixture modeling over TED-only baselines. We also took advantage of phrase and reordering table fill-up models for further domain adaptation that additionally compresses the size of the translation system.

In Arabic-English, we used early distortion cost and incorporated a hybrid word/class language model to adapt to the style of talks, while for Germanic languages, we explored the effects of various compound splitting techniques. For Turkish-English, we compared several approaches to morphological segmentation and used a hierarchical SMT system.

9. Acknowledgements

This work was partially supported by TOSCA-MP project (IST-287532) and the EU-BRIDGE project (IST-287658), which are both funded by the European Commission under the Seventh Framework Programme for Research and Technological Development.

10. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, December 2012.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] R. C. Moore and W. Lewis, "Intelligent selection of language model training data," in *ACL (Short Papers)*, 2010, pp. 220–224.

[10] 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.

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

[12] P. Clarkson and A. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, Munich, Germany, 1997, pp. 799–802.

[13] J. Clark, C. Dyer, A. Lavie, and N. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proceedings of the Association for Computational Lingustics*, ser. ACL 2011. Portland, Oregon, USA: Association for Computational Linguistics, 2011, available at http://www.cs.cmu.edu/ jhclark/pubs/significance.pdf.

[14] R. C. Moore and C. Quirk, "Faster beam-search decoding for phrasal statistical machine translation," in *In Proceedings of MT Summit XI*, 2007.

[15] A. Bisazza and M. Federico, "Cutting the long tail: Hybrid language models for translation style adaptation," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, April 2012, pp. 439–448.

[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] K. Oflazer, "Two-level description of Turkish morphology," *Literary and Linguistic Computing*, vol. 9, no. 2, pp. 137–148, 1994.

[18] T. G. H. Sak and M. Saraçlar, "Morphological disambiguation of Turkish text with perceptron algorithm," in *Proc. of CICLing*, 2007, pp. 107–118.

[19] A. Bisazza and M. Federico, "Morphological preprocessing for turkish to english statistical machine translation," in *International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, 2009.

[20] A. Bisazza, I. Klasinas, M. Cettolo, and M. Federico, "FBK @ IWSLT 2010," in *International Workshop on Spoken Language Translation (IWSLT)*, Paris, France, 2010.

[21] M. Creutz and K. Lagus, "Inducing the morphological lexicon of a natural language from unannotated text," in *International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

[22] H. Erdoğan, O. Büyük, and K. Oflazer, "Incorporating language constraints in sub-word based speech recognition," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, 2005, pp. 98–103.

[23] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 263–270.

[24] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*, 2003.

[25] K. Macherey, A. Dai, D. Talbot, A. Popat, and F. Och, "Language-independent compound splitting with morphological operations," in *Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics (ACL)*. Portland, USA: Association for Computational Linguistics, 2011.

[26] D. Varga, , L. Nmeth, P. Halacsy, A. Kornai, V. Tron, and V. Nagy, "Parallel corpora for medium density languages," in *Proceedings of the RANLP 2005*, 2005, pp. 590–596.