FBK's Machine Translation Systems for the IWSLT 2013 Evaluation Campaign

Nicola Bertoldi[1], *M. Amin Farajian*[1,2], *Prashant Mathur*[1,2], *Nicholas Ruiz*[1,2], *Marcello Federico*[1]

[1]Fondazione Bruno Kessler
HLT Unit
Via Sommarive, 18
38123 Trento, TN, Italy

[2]University of Trento
ICT Doctoral School
Via Sommarive, 5
38123 Trento, TN, Italy

Abstract

This paper describes the systems submitted by FBK for the MT track of IWSLT 2013. We participated in the English-French as well as the bidirectional Persian-English translation tasks. We report substantial improvements in our English-French systems over last year's baselines, largely due to improved techniques of combining translation and language models. For our Persian-English and English-Persian systems, we observe substantive improvements over baselines submitted by the workshop organizers, due to enhanced language-specific text normalization and the creation of a large monolingual news corpus in Persian.

1. Introduction

FBK's machine translation activities in the IWSLT 2013 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 describe our participation in the Machine Translation translation tasks in the official English-French as well as the optional English-Persian and Persian-English languages. These tasks entail translating subtitles transcribed and translated by the TED community.

We begin with an overview of the domain adaptation techniques used by each of our language pair experiments in Section 2: namely, data filtering and methods to combine translation models, reordering models, and language models from multiple corpora, respectively. In Section 3, we describe several experiments in the English-French translation task. In Section 4, we describe our first efforts at translated to and from English and Persian, a language pair with few parallel resources available. We introduce our efforts to collect and preprocess Perian corpora to improve the quality of Persian translation and show significant improvements over the state of the art. In Section 5 we summarize our findings.

For all language pairs, we set up a standard phrase-based system using the Moses toolkit [2]. We construct a statistical log-linear models including domain-adapted phrase translation and hierarchical reordering models [3, 4, 5], one or more target language models (LM), as well as distortion, word, and phrase penalties.

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

2. Domain adaptation techniques

In this section, we summarize several well-known techniques for domain adaptation we applied to build high-performance models for our SMT submissions.

2.1. Data selection

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.

To this purpose, we follow the procedure described in [6], which adapts the cross-entropy difference scoring technique introduced by [7] toward bitext data selection. First, all sentence pairs of the out-of-domain corpus are associated with a source- and target-side scores, each computed as the basic technique proposes for the corresponding monolingual scenarios, using the in-domain (TED) data as a seed and LMs of order 3. Then, the sentences are sorted according to the sum of these two scores. Finally, the optimal split between useful and useless sentences is found by minimizing the source-side perplexity of a development set on growing percentages of the sorted corpus. In our experiments, dev2010 and tst2010 are concatenated and used as the filtering development set.

2.2. Translation model combination

Three methods are applied in our submissions to combine the TM built on the available parallel training corpora: namely, fill-up [8, 9], back-off, and interpolation.

2.2.1. Fill-up and Back-off

In the fill-up approach, out-of-domain phrase pairs that do not appear in an in-domain (TED) phrase table are added, along with their scores – effectively filling the in-domain table with additional phrase translation options. The fill-up process is performed in a cascaded order, first filling in miss-

ing phrases from the corpora that are closest in domain to TED. Moreover, out-of-domain phrase pairs with more than four source tokens are pruned.

Following [8, 9] the fill-up approach adds k-1 provenance binary features to weight the importance of out-of-domain data, where k is the number of phrase tables to combine. A similar back-off approach performs the fill-up technique, but does not add any provenance binary features.

2.2.2. Linear interpolation

A common approach for building multi-model is through the linear interpolation of component models. Various approaches have been suggested for computing the coefficients of the interpolated model, the most recent being perplexity minimization described in [10] where the perplexity of each component translation model is minimized on the parallel development set. However, the mixing coefficients can be separately computed by several other techniques. In this paper, instead of calculating translation model perplexity we calculate language model perplexity on target side development set. After minimizing perplexity we get the interpolation weights which we then use as mixing coefficients for component translation models.

2.3. Reordering model combination

All techniques available for combining the TMs can be applied straightfowardly to combine the RMs. The only difference regards the fill-up technique: the additional binary feature is discarded, since it is already present in the corresponding filled-up TM. Hence, a filled-up RM is exactly the same as a backed-off RM.

2.4. Language model combination

Language models are built from the monolingual training data, as well as the target language of the parallel data. As the corpora available in the IWSLT evaluation come from a number of sources, we apply several methods to combine the LMs built on the available target language training corpora, rather than concatenating the data.

2.4.1. Mixture

Monolingual subcorpora can be combined into one mixture language model [11] by means of the IRSTLM toolkit [12]. The optimization of the internal mixture weights is achieved through a cross-validation approach on the same training data; hence no external development set is required. The mixture LM type can be loaded by Moses as any other LM type.

2.4.2. Linear interpolation

This technique, provided by the IRSTLM toolkit, consists in the linear interpolation of the n-gram probabilities from all component LMs. The optimal interpolation weights are computed by the EM algorithm which minimizes the perplexity on a given held-out development sample. The IRSTLM toolkit provides an interface that enables Moses to compute n-gram probabilities from interpolated LMs.

2.4.3. Log-linear interpolation

This technique, provided directly within the Moses toolkit, consists in the log-linear interpolation of the n-gram probabilities from all component LMs. The weight optimization is performed during the tuning of all Moses features.

3. English-French system

Our English-French systems are built upon a standard phrase-based system using the Moses toolkit [2], exploiting a huge amount of English-French bitexts and monolingual French training data. Each system features a statistical log-linear model including one phrase translation model [9] and one lexicalized reordering model, multiple French language models (LMs), as well as distortion, word, and phrase penalties.

The training data are composed from some of the corpora allowed by the IWSLT Evaluation Campaign organizers. As parallel data the following corpora were taken into account: Web Inventory of Transcribed and Translated Talks (version 2013-01) (TED) [13], 10^9-French-English (version 2) (Giga), English-French Europarl (version 7) (EP), Common Crawl (CC), MultiUN (UN), and the News Commentary (News) corpus as distributed by the organizers of the Workshop of Machine Translation (WMT). As monolingual data we use the entire monolingual news corpora (Full) distributed by WMT organizers for language model training. All texts were processed according to the language specific tokenization provided by Moses toolkit and kept case-sensitive. Statistics of the training corpora are reported in Table 1.

	unselected			selected		
Corpus	**Segm**	**En Words**	**Fr Words**	**Segm**	**En Words**	**Fr Words**
TED	155.5K	3.1M	3.2M	155.5K	3.1M	3.2M
Giga	22.5M	662.8M	774.7M	1.1M	23.8M	26.9M
UN	12.9M	361.6M	413.1M	257.7K	5.1M	5.6M
CC	3.2M	80.7M	88.0M	973.2K	23.2M	25.2M
EP	2.0M	55.6M	60.0M	240.9K	5.1M	4.8M
News	170.2K	4.4M	5.0M	51.1K	1.1M	1.3M
Full	84.0M	na	2.4T	na		

Table 1: Statistics of the parallel and monolingual data exploited for training our English-French systems. For the parallel data, statistics before and after data selection are reported. Symbols "T", "M" and "K" stand for 10^9, 10^6 and 10^3, respectively.

In order to focus the models toward a TED-specific domain and genre and to reduce the model size, data selection by means of the IRSTLM toolkit [12] is performed on the English-French bitexts, using the TED training data as in-domain data. Different amounts of data are selected from

each of the available out-of-domain corpora; statistics are reported in Table 1. A detailed description of the data selection procedure is provided in Section 2.1.

We construct five systems which exploit the training data in different ways to construct the component models. Details for these systems are provided in Section 3.1.

Most system parameters are kept fixed to allow a better comparison among the systems. Word alignments are computed by means of MGIZA++ on case-insensitive parallel texts to reduce data sparseness; casing information is re-introduced in order to estimate case-sensitive models, unless otherwise specified in the particular experiment. In all systems the maximum phrase length is set to 7 and the distortion limit is set to the default value of 6. We train 5-gram LMs with IRSTLM toolkit [12] in most cases; in other cases, KenLM [14] is used. Each language model is smoothed via the improved Kneser-Ney technique. Singleton n-grams of order three or higher are pruned.

The weights of the log-linear combination are optimized either via minimum error rate training (MERT) [15] or the Margin Infused Relaxed Algorithm (MIRA) [16, 17] on dev2010.

3.1. English-French submissions

As described in Section 3, we submit five systems which differ in the exploitation of the training data for the creation of TM, RM and LMs. We evaluate the performance of each system in Table 2 and use the results on tst2010 to select our primary submission. In our Primary, Contrastive 1, and Contrastive 2 systems, the dev2010 and tst2010 data are added to the TED training data after optimizing each system's feature weights, before evaluating their performances on the 2011, 2012, and 2013 test sets.

3.1.1. Primary

A backed-off TM is created combining a primary TM trained on TED training data (TED-TM) and a background TM trained on the selected training data (Slct-TM). The RM is constructed in a similar manner. A log-linear combination of two LMs is employed. The first LM is a mixture estimated from the in-domain TED training data (TED-LM) and the out-of-domain data-selected training data (Slct-LM). Additionally, a second Full-LM is estimated from the entire French monolingual corpora. Minimum Bayes Risk [18] (MBR) decoding technique, provided by Moses, is also exploited. Feature weights are averaged over three MERT optimizations.

3.1.2. Contrastive 1

This system replaces the backed-off TM of the primary system with a filled-up TM that exploits the same component TMs. Moreover, the MBR decoding technique is not applied. The feature weights are newly estimated averaging three distinct MERT optimizations.

3.1.3. Contrastive 2

This system aims at enhancing the primary system by further focusing its models to each specific talk that comprises the test set. Using the same optimized feature weights, we construct talk-specific translation, reordering, and language models and insert them with highest priority in their respective back-off and mixture models.

Given a talk to translate, we perform the data selection procedure described in Section 2.1, using the source text of the talk as seed data to extract the most similar portion from the data-selected parallel training data. Unlike the training phase, this selection is based on the English monolingual score only and a fixed amount of parallel data (about 3.5M English running words) were extracted.

Like the primary system, MBR decoding is applied. It is worth highlighting that this system is actually a collection of talk-specific instances working on their corresponding talk.

3.1.4. Enhanced Contrastive 2

In the post-evaluation activity, we performed an ad-hoc tuning of the system weights. For each talk of tst2010, we search for the optimal weights of the corresponding talk-specific system with our standard MERT procedure; then, all talk-specific weight sets are averaged and exploited for running the system over the official tst2011-2013. We also test this enhanced system on tst2010 in a fair manner: when translating a talk we exclude the corresponding set of optimal weight during the averaging action.

3.1.5. Contrastive 3

Following [10], the corpus specific TMs and RMs are combined according to the linear interpolation technique, but a different procedure is performed to find the mixing coefficients of the linear-interpolated TM and RM. A linear-interpolated LM is created by combining the corpus-specific LMs and its mixing coefficients are optimized by minimizing the perplexity on dev2010 target side using Expectation-Maximization by means of the IRSTLM toolkit. These interpolation weights are utilized as mixing coefficients for the linear-interpolated TM and RM. In this system we employ all LMs, estimated on the each of the 6 different domains, and the Full-LM combined in a log-linear fashion.

The system applies MBR decoding and case-insensitive models; therefore, a re-casing module estimated on the training data is attached to the translation system.

The whole set of the Moses features weights are optimized running the MIRA algorithm once.

3.1.6. Contrastive 4

This system differs from contrastive 3 only in the number of employed LMs; rather than using a log-linear combination of seven LMs, it utilizes only two: namely, TED-LM and Full-LM.

3.2. English-French results

Performance in terms of case-sensitive BLEU and TER of our primary (*P*) and contrastive (*C*) systems are reported in Table 2 and are compared to a simple TED baseline[2] (*B*). This baseline relies on TED training data only for the estimation of its TM, RM, and LM; the second Full LM is employed as well.

Figures referred to tst2010 were computed in-house, while those for tst2011-2013 are the official results provided by the organizers. As the official evaluation uses a slightly different text normalization procedure, the absolute scores are not directly comparable between different test sets; nevertheless, the relative difference among the systems are reliable.

In the result tables, the $^{\blacktriangledown}$ and $^{\triangledown}$ symbols beside the BLEU and TER scores indicate that the corresponding system performs significantly worse than the primary system with p-values not larger than 0.01 and 0.10, respectively. This annotation regards tst2010 only, for which the reference translations are available and hence the significance test can be performed.

	BLEU				TER			
	tst_{10}	**tst_{11}**	**tst_{12}**	**tst_{13}**	**tst_{10}**	**tst_{11}**	**tst_{12}**	**tst_{13}**
P	**34.11**	**38.41**	**39.51**	**37.69**	**0.472**	**0.420**	**0.406**	**0.441**
C_1	33.79$^{\triangledown}$	37.84	39.44	37.60	0.478$^{\blacktriangledown}$	0.426	0.409	0.441
C_2	31.90$^{\blacktriangledown}$	35.16	36.60	35.17	0.489$^{\blacktriangledown}$	0.443	0.429	0.458
C_3	34.03	28.99	29.69	26.36	0.479$^{\blacktriangledown}$	0.511	0.496	0.550
C_4	33.61$^{\triangledown}$	28.83	29.36	26.35	0.480$^{\blacktriangledown}$	0.511	0.498	0.548

Table 2: Results of the official English-French submissions evaluated on the IWSLT TED test sets. Symbols $^{\blacktriangledown}$ and $^{\triangledown}$ near to BLEU and TER scores on tst_{10} indicate that the system performs significantly worse than the primary system with p-values not larger than 0.01 and 0.10, respectively.

We can draw out some comments from the analysis of the official results. The primary system consistently outperforms the contrastive systems, and differences in scores are somehow kept constant. The improvement over the reference baseline system (shown in Table 3) is strongly significant, proving the effectiveness of the data selection approach applied

The low scores achieved by C_3 and C_4 on the 2011-2013 test sets are due to a misconfiguration of these systems when applied to the official data sets. After the official evaluation we translated the test sets with the corrected systems (C_3^* and C_4^*), and asked the organizers to re-evaluate them. New results are reported in Table 3. Scores for tst_{11}, tst_{12}, and tst_{13} have been computed by means of a different evaluation script; hence, figures in Tables 2 and 3 are not directly comparable.

On tst2010, all systems, but C_2, achieve very similar results in terms of both BLEU and TER. This is somehow expected, because the systems have very similar configurations. In terms of BLEU, a statistical test shows a slightly significant difference with respect to *P* only for C_1 and C_4, and only at p-value of 0.10. Instead, the differences in terms of TER are always significant.

Interestingly, from the results of C_3^* and C_4^*, we observe that the log linear combination of 6 language models does not improve the performance of the system, but instead it has negative effects on `tst2011` and `tst2013`. Use of out-domain language models diverge the "virtual domain" of interpolated TM and RM away from TED domain. The main difference between C_4^* and *P* is the way of combining TMs and RMs. *P* uses the back-off approach while C_4^* uses linear interpolation. This basically shows that back-off performs better than the linear interpolation technique for TED-talks data.

System C_2 is statistically worse than *P*. Our preliminary analysis showed that this system produced translation outputs about 4% shorter than *P*. Our feeling is that this is due to the exploitation of log-linear weights not specifically estimated for the talk-specific system. In order to confirm our conjecture, we translated the test sets with the enhanced system (C_2^*) described in Section 3.1, and its performance are reported in Table 3. It outperforms the primary system in terms of BLEU, but the differences are not significant, at least on tst2010. Instead, its performance in terms of TER are worse than those of the primary system; this is probably due to the fact that weight optimization aims at improving only the BLEU metric. A more balanced improvement could be achieved by tuning over several metrics.

	BLEU				TER			
	tst_{10}	**tst_{11}**	**tst_{12}**	**tst_{13}**	**tst_{10}**	**tst_{11}**	**tst_{12}**	**tst_{13}**
B	32.43$^{\blacktriangledown}$	35.77	36.95	34.56	0.489$^{\blacktriangledown}$	0.426	0.413	0.457
P	34.11	37.53	38.83	37.10	0.472	0.412	0.397	0.437
C_1	33.79$^{\triangledown}$	37.05	38.70	37.05	0.478$^{\blacktriangledown}$	0.418	0.401	0.433
C_2	31.90$^{\blacktriangledown}$	34.42	36.08	34.76	0.489$^{\blacktriangledown}$	0.436	0.421	0.450
C_2^*	34.28	38.72	39.80	37.68	0.486$^{\blacktriangledown}$	0.413	0.407	0.444
C_3^*	34.03	36.95	38.40	36.26	0.479$^{\blacktriangledown}$	0.423	0.405	0.449
C_4^*	33.61$^{\triangledown}$	37.28	38.14	36.42	0.480$^{\blacktriangledown}$	0.423	0.407	0.447

Table 3: Results of official and unofficial English-French submissions evaluated on the IWSLT TED test sets. C_2^*, C_3^*, and C_4^* are unofficial revised submissions. Scores for tst_{11}, tst_{12}, and tst_{13} have been computed by the organizers by means of an evaluation script partially different from the official one. Symbols $^{\blacktriangledown}$ and $^{\triangledown}$ near to BLEU and TER scores on tst_{10} indicate that the system performs significantly worse than the primary system with p-values not larger than 0.01 and 0.10, respectively.

4. English-Persian systems

The Persian-English (Fa[3]-En) and English-Persian (En-Fa) systems are built using similar configurations to our English-French system, described in Section 3. To relax the problem of token inconsistencies in Persian documents, we devel-

[2]System *B* was not submitted for the official evaluation, and therefore no results for tst2011-2013 are available.

[3]According to ISO 639-1 (*Codes for the representation of names of languages*), "Fa" is used as the abbreviation of Persian.

oped a Persian text normalizer that yields consistently better translation than the unnormalized text. Furthermore, to have a more precise Persian LM, we created a large Persian monolingual corpus by crawling feeds from several online news agencies. We show that the combination of specialized text normalization and a large LM trained on additional Persian data provides substantitive improvements over previous baselines.

4.1. Persian Text Normalization and Tokenization

Although there are some electronic standards for writing Persian, they are not uniformly followed by writers and software tools. These inconsistencies are observed in all existing textual resources, which cause many problems in natural language processing tasks. Several problems that commonly result in separate tokens for redundant types are described below.

Different character sets may be used for the same letter. Their appearance is virtually the same but different encodings exist for the characters. YEH(ي) and KAF (ك) are the best known cases in this category. On the other hand, some authors prefer to use imported letters from Arabic (e.g. أ) for writing the words borrowed from Arabic (رأى), while others use Persian letters (رای).

Diacritics are not typically written in the standard Persian text, but some authors decide to use them to reduce the ambiguty of the words. Although this makes the text more clear and understandable to the reader, not all authors use diacritic marks. Without proper preprocessing, the text processing system cannot classify different instances of the same word into one class.

Different word forms. This problem is mostly due to the word boundary ambiguation and different ways of putting space between different parts of words. For example, the word می روم (I am going) can be written in any of the following forms: می روم, می‌روم, and میروم. In the first and second forms, the prefix می and verb روم are separated using space and zero-width non-joiner (ZWNJ) characters, respectively; while in the last case, the prefix is attached to the verb.

To relax the problem of token inconsistency, we developed a Persian text normalizer and applied it on all of the Persian texts used in the experiments. This normalizer is published by the organizers and used to normalize all MT outputs and references before evaluating the systems in the English-Persian language pair tracks. A version of this tool was released for use with the IWSLT 2013 shared task. An enhanced version will be publicly available in the near future.

To measure the usefulness of the normalizer we develop two baseline systems using the normalized and non-normalized training data, and evaluate their translation quality in Table 4. The results show significant improvements in the final quality of the systems in both directions (1.5+ in BLEU scores and 2.7+ in TER). Furthurmore, comparing the vocabulary size of the normalized and non-normalized training corpus, shows more than 11 percent reduction in the number of unique words.

	BLEU		TER	
Metric	**Fa-En**	**En-Fa**	**Fa-En**	**En-Fa**
Baseline	12.47	9.13	0.734	0.758
Normalized	13.94	10.70	0.706	0.725

Table 4: Comparing the results of the normalized and unnormalized baselines on the IWSLT TED test set 2010.

4.2. Data Preparation

The data provided by the organizers for the Persian-English task is only the TED corpus; no additional parallel or monolingual corpora are provided for Persian. Though there are some other publicly available parallel corpora (namely, TEP [19], and PEN [20]), our initial experiments showed that using these corpora do not improve the baseline. Therefore, we decided not to use them in our submissions.

Regarding monolingual corpora, the Hamshahri corpus [21], used widely used in different Persian text processing tasks, has inconsistent sentence boundaries in such a way that in many cases one sentence is split into several lines, with no boundary markers in the corpus to capture the complete sentence.

Since this affects the language model creation and decreases the accuracy of the LM, we decided to create our own large Persian monolingual corpus with proper sentence boundaries. To create this corpus we extract texts from the archives of more than 20 online news agencies, mainly located in Iran. We extract the *body* of the news stories, as well as the *title*, *publish date*, and the *genre*, if available. The documents smaller than 1K are filtered out in this step. We normalize the documents using the aforementioned normalizer. The statistics of the corpus are presented in Table 5. This corpus will be publicly released at a future date.

		Tokens		Types	
Corpus	**Segm**	**English**	**Persian**	**English**	**Persian**
TED	77.1K	1.5M	1.7M	16.4K	20.8K
FBK	11.2M	na	309.2M	na	536.2K
FBK-slct	3.6M	na	50.1M	na	213K

Table 5: Statistics of the parallel and monolingual data exploited for training purpose in the English-Persian and Persian-English systems. Symbols "M" and "K" stand for 10^6 and 10^3, respectively. "FBK-slct" refers to the data selected portion of our internal Persian monolingual corpus.

For our Persian-English MT submission, 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. For the English-Persian task we used the TED training data (Persian side) and the monolin-

	BLEU				TER			
	$\mathbf{tst}_{10}$	$\mathbf{tst}_{11}$	$\mathbf{tst}_{12}$	$\mathbf{tst}_{13}$	$\mathbf{tst}_{10}$	$\mathbf{tst}_{11}$	$\mathbf{tst}_{12}$	$\mathbf{tst}_{13}$
B	12.47	16.39	12.80	12.49	0.734	0.678	0.88	0.876
P	14.62	18.85	14.40	14.32	0.703	0.664	0.861	0.858
C_1	–	–	–	14.47	–	–	–	0.858

Table 6: Results of submitted Persian-English runs evaluated on the IWSLT TED test sets.

	BLEU				TER			
	$\mathbf{tst}_{10}$	$\mathbf{tst}_{11}$	$\mathbf{tst}_{12}$	$\mathbf{tst}_{13}$	$\mathbf{tst}_{10}$	$\mathbf{tst}_{11}$	$\mathbf{tst}_{12}$	$\mathbf{tst}_{13}$
B	9.13	11.57	9.67	8.93	0.758	0.718	0.741	0.727
P	11.55	12.55	10.94	10.12	0.723	0.701	0.727	0.716
C_1	–	–	–	10.32	–	–	–	0.715

Table 7: Results of submitted English-Persian runs evaluated on the IWSLT TED test sets.

gual corpus described earlier.

4.3. English-Persian submissions

For both English-Persian and Persian-English tasks, we submitted a primary and a contrastive systems, which are briefly described in the following.

4.3.1. Primary

Our primary system uses the text normalization approach described in Section 4.1. For both the English-Persian and Persian-English submissions a TM is trained on TED training data, using similar configurations to our English-French systems, described in Section 3. For the Persian-English submission a log-linear combination of two LMs is employed. The primary LM is a 5-gram LM, trained on TED training data (English side), while the second LM is a 5-gram mixture LM consisting of TED data and the out-of-domain data-selected training data.

In the English-Persian direction the log-linear combination of LMs consist of three 5-gram LMs, trained on TED data, data selected from FBK Persian monolingual corpus, and whole FBK Persian monolingual corpus, respectively. As in our primary English-French submission, Minimum Bayes Risk decoding is exploited. Feature weights are optimized via Margin Infused Relaxed Algorithm (MIRA) on dev2010.

4.3.2. Contrastive

As mentioned earlier, the English-Persian language pair has few bitexts available for constructing a translation model. To measure the effects of adding additional in-domain corpora on translation quality, we augment the translation and reordering models with tst2011 and tst2012 and evaluate the results on tst2013 while retaining the log-linear weights of the original models.

4.4. English-Persian results

Our primary (*P*) and constrastive (*C*) results for Persian-English and English-Persian are reported in Tables 6 and 7, respectively. We compare the performance of our systems against a simple baseline (*B*), trained on the unnormalized TED data only. Scores on tst2010 clearly prove that our primary system highly outperforms the baseline.

The small amount of additional training data exploited in the contrastive system only gives a slight improvement in BLEU.

Long distance reorderings and the morphological richness of Persian are the two major problems in Persian-English SMT systems. On the other hand, hierarchical models are known to outperform the phrase-based systems for language pairs with differing word orders or long-distance reorderings. Our primary experiments in using hierarchical models for this langauge pair do not outperform the phrase-based baseline system, however. We will investigate this in more detail in future work.

One technique to overcome data sparsity due to morphological inflections is to perform unsupervised segmentation [22] and using the root forms for word alignment. However, in preliminary experiments we did not observe improvements over a baseline that only considers the surface form. One reason for this behavior may be due the fact that the suffixes they carry meaning that is lost during word alignment, which subsequently affects the quality of the extracted phrases. In the future we plan to try other morphological analysis strategies that better model the characteristics of Persian.

5. Conclusion

We presented the MT systems with which we participated in the IWSLT 2013 TED MT Evaluation Campaign. Our English-French systems benefited most from a "back-off" combination of in-domain and out-of-domain translation models, as well as a log-linear combination of two language model flavors: one which combines corpus-specific language models in a mixture model, and the other that concatenates all corpora and generates a gigantic LM.

Our English-Persian and Persian-English systems showed substantial improvements over a baseline provided by the workshop organizers, largely from improving the normalization and tokenization of Persian texts, as well as acquiring a large monolingual Persian news crawl corpus.

6. Acknowledgments

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

7. References

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, M. Federico, "Report on the 10th IWSLT Evaluation

Campaign," in *Proc. of the International Workshop on Spoken Language Translation*, December 2013.

[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] M. Cettolo, C. Servan, N. Bertoldi, M. Federico, L. Barrault, and H. Schwenk, "Issues in Incremental Adaptation of Statistical MT from Human Post-edits," in *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice (WPTP-2)*, Nice, France, September 2013.

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

[8] 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.

[9] 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, pp. 136–143.

[10] R. Sennrich, "Perplexity minimization for translation model domain adaptation in statistical machine translation," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association For Computational Linguistics, 2012, pp. 539–549.

[11] M. Federico and R. De Mori, "Language modelling," in *Spoken Dialogues with Computers*, R. D. Mori, Ed. London, UK: Academy Press, 1998, ch. 7, pp. 199–230.

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

[13] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012.

[14] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July 2011, pp. 187–197.

[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.

[16] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 551–585, 2006.

[17] T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki, "Online large-margin training for statistical machine translation," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 764–773.

[18] S. Kumar and W. Byrne, "Minimum bayes-risk decoding 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.

[19] M. Pilevar, H. Faili, and A. Pilevar, "Tep: Tehran english-persian parallel corpus," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2011, vol. 6609, pp. 68–79.

[20] M. A. Farajian, "Pen: Parallel english-persian news corpus," in *Proceedings of 2011 International Conference on Artificial Intelligence (ICAI'11)*, Las Vegas, NV, 2011.

[21] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard persian text collection," *Know.-Based Syst.*, vol. 22, no. 5, pp. 382–387, July 2009.

[22] 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.