FBK's Machine Translation and Speech Translation Systems for the IWSLT 2014 Evaluation Campaign

Nicola Bertoldi[1], *Prashant Mathur*[1,2], *Nicholas Ruiz*[1,2], *Marcello Federico*[1]

[1]Fondazione Bruno Kessler
Human Language Technologies
Trento, Italy

[2]University of Trento
ICT Doctoral School
Trento, Italy

Abstract

This paper describes the systems submitted by FBK for the MT and SLT tracks of IWSLT 2014. We participated in the English-French and German-English machine translation tasks, as well as the English-French speech translation task. We report improvements in our English-French MT systems over last year's baselines, largely due to improved techniques of combining translation and language models, and using huge language models. For our German-English system, we experimented with a novel domain adaptation technique. For both language pairs we also applied a novel word triggers based model which shows slight improvements on English-French and German-English systems. Our English-French SLT system utilizes MT-based punctuation insertion, recasing, and ASR-like synthesized MT training data.

1. Introduction

FBK's machine translation activities in the IWSLT 2014 Evaluation Campaign 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 English-French and German-English Machine Translation tasks as well as in the English-French Spoken Language Translation task.

After a brief introduction to the baseline MT system in Section 2 employed for all tasks, in Section 3 we overview the data selection techniques used to extract TED-related data from the available huge and generic monolingual and bilingual corpora. Then, in Section 4 we describe the methods applied to combine translation models, reordering models, and language models trained on multiple corpora. Sections 5-7 give details about the actual MT and SLT systems built for evaluation task.

2. Baseline SMT system

All our task-specific systems rely on the well-known and state-of-the-art phrase-based Moses toolkit [1]; and exploit the huge amount of parallel and monolingual training data provided by the organizers. Our common baseline system features a statistical log-linear model including a phrase-based translation model (TM), a lexicalized phrase-based reordering models (RM), one or more language models (LMs), as well as distortion, word and phrase penalties.

Tuning of the baseline system is performed on tst2010 by optimizing BLEU using Minimum Error Rate Training [2]. However, all available development data sets, namely dev2010 and tst2010-2012, are included in the in-domain training data to build the systems actually employed for the 2014 evaluation campaign. The task-specific systems differ in the way training data are processed and filtered, and how the models are trained and combined.

3. Data Filtering

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 [3], implementing the bilingual cross-entropy difference [4], i.e. an adaptation of the cross-entropy difference scoring technique introduced by [5] toward bitext data selection, by means of XenC toolkit [6].

First, all sentence pairs of the out-of-domain corpus are associated with source- and target-side scores, each of which are computed as the basic technique proposes for the corresponding monolingual scenarios. We use the in-domain (TED) data as a seed and LMs of order 2.[2] 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.

4. Domain Adaptation

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

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

[2] This small LM order permits a very fast computation of the scores, without losing performance.

4.1. Translation model combination

Three methods are applied in our submissions to combine the TM built on the available parallel training corpora: namely, fill-up [7, 8], back-off, and interpolation.

4.1.1. Fill-up

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 missing 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 [7, 8] 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.

4.1.2. Back-off

The back-off approach works similarly to the fill-up technique, but does not add any provenance binary features.

4.1.3. Linear interpolation

Linear interpolation of component models is a widely used approach for building a domain adapted multi-model. Approaches such as using monolingual data or pairwise ranking optimization to set interpolation weights [9, 10], perplexity minimization [11], and combining lemmatized and non-lemmatized models [12] have been used in the past for improved domain adaptation. In this paper, we leverage a recent work of [13] which exploits the use of source-side of the parallel in-domain corpus for domain adaptation. This approach calculates a similarity score (known as BLEU-PT) for each of the out-domain translation models on the source in-domain data. We use these similarity scores and further normalize them by the number of phrases seen in each of the corresponding out-domain phrase tables. These normalized scores are then used as linear interpolation coefficients.

In this paper, we perform linear interpolation of out-of-domain models which results in one translation model. The in-domain translation model is then filled-up with the aforementioned interpolated out-domain translation model giving us a single domain adapted model.

4.2. Reordering model combination

All techniques available for combining the TMs can be applied straightforwardly 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.

4.3. 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.

4.3.1. Mixture

Monolingual subcorpora can be combined into one mixture language model [14] by means of the IRSTLM toolkit [15]. 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.

4.3.2. 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.

4.4. Factored Trigger Models

Cross-lingual lexical triggers have been already studied in natural language processing [16] and in machine translation [17]. The latter defined cross lingual triggers as a setup of a trigger word (f_i) in the source language sentence, triggering a number of words ($e_0, e_1, \ldots, e_n$) in the target language sentence. For each trigger source word f_i, we calculate pointwise mutual information (PMI) between that word and the target triggered words (e_j) as shown in the equation 1.

$$PMI_{lex}(f_i, e_j) = log \frac{P(f_i, e_j)}{P(f_i) \cdot P(e_j)} \quad (1)$$

In this paper, we extend these lexical triggers with additional factors such as POS tags and lemmas. Similar to computing the PMI for lexical triggers we compute corresponding PMIs for the POS tags and lemmas of the trigger and the triggered words in question. This is shown in the following equations:

$$PMI_{pos}(f_i, e_j) = log \frac{P(POS(f_i), POS(e_j))}{P(POS(f_i)) \cdot P(POS(e_j))} \quad (2)$$

$$PMI_{lemma}(f_i, e_j) = log \frac{P(LEM(f_i), LEM(e_j))}{P(LEM(f_i)) \cdot P(LEM(e_j))} \quad (3)$$

where $POS(x)$ is the part of speech tag of the word x and $LEM(x)$ is the lemma of the word x. These PMI are computed for all word pairs and then normalized over the whole parallel corpus. In the end, a factored trigger model (henceforth, FTM) contains three different features for each of the source/target word pair.

At decoding time, when a phrase-based machine translation system requests feature values from the FTM for a phrase pair $(f_{i,\ldots,j}, e_{k,\ldots,l})$, it returns the average sum of all the feature values for all word pairs possible within the phrase pair. Mathematically, it can be denoted as the following:

$$FTM_{lex}(f_{i,\ldots,j}, e_{k,\ldots,l}) = \sum_{z=i}^{j} \sum_{y=k}^{l} PMI_{lex}(f_z, e_y). \quad (4)$$

Similarly, POS and Lemma features are also calculated at the run-time and fed directly to the decoder providing a seamless integration of factored trigger model in a phrase based machine translation system. This integration also allows us to use any tuning algorithm (e.g. MERT, MIRA) easily.

5. English-French MT task

In order to adapt the English-French MT system to the TED-specific domain and genre, as well as to reduce the size of the models, data selection (see Section 3) is carried out on several parallel English-French corpora provided by the organizers, namely Europarl, CommonCrawl, UN, News Commentary, News Crawl, and Giga, and using the whole WIT[3] [18] training corpus as in-domain data.

Different amount of texts were selected from each corpus ranging from 2% to 30%, which are concatenated together to build one large parallel corpus containing 2.6M sentences for a total of 57M English and 63M French running words.

The system for FBK primary submission is built as follows. Two TMs and two RMs are trained independently on the parallel in-domain and selected data, using the standard Moses procedure and MGIZA++ toolkit [19] for word-alignment; TMs and RMs were combined using the back-off technique (for both TM and RM), taking WIT[3] as primary component, for a total of 168M phrase pairs.

The French side of the in-domain and selected data are also employed to estimate a 2-component mixture language model (see Section 4.3). A second huge French LM is estimated as an 8-component mixture on all permitted monolingual French data: namely, the target side of the parallel training corpora,[3] consisting of about 1.4G running words. Both LMs have order 5 and are smoothed by means of the interpolated Improved Kneser-Ney method [20]; they include 57M and 661M 5-grams, respectively. Finally, the three additional features provided by the factored trigger model (see Section 4.4) are included in the log-linear combination.

Minimum Bayes Risk (MBR) [21] decoding is applied with its default values.

As already mentioned in Section 2, all available development data sets, namely dev2010 and test2010-2012, are included in the in-domain training data to build the primary system.

[3]The monolingual French Gigaword Third Edition replaces the French side of the parallel Giga English-French corpus employed in the TM and RM model training.

In order to evaluate the contribution of the individual components of the FBK system, we submitted several contrastive runs.

- contrastive-7: derived from primary system, this system does not exploit the factored trigger model;
- contrastive-6: derived from contrastive-7, this system exploits the stack decoding instead of the MBR decoding;
- contrastive-5: derived from contrastive-6, this system does not exploit the huge French LM.

Moreover, we submitted 4 runs (contrastive 1-4) which differ from contrastive 5-7 and the primary run just in one aspect: contrastive 1-4 do not include the development data sets in the training data. The aim was to measure the impact of a limited amount of additional TED talks on the translation quality.

Finally a ninth run (contrastive-9) was submitted with a system built on top of the primary, which tests the assumption made during translation modeling that each of the features in the translation model are independent from one another. Generalized linear models can be constructed in a manner that models interactions between predictors (e.g. [22]). As a preliminary experiment, we test for interactions between the forward and backward phrase probabilities in our phrase table, expressed as a multiplication between the log probabilities.

Several observations can be drawn from the analysis of the figures reported in Table 1, also supported from preliminary experiments performed during the development phase.[4]

- The biggest performance improvement is due to the use of the large French LMs.
- MBR decoding gives a small but consistent boost in quality with respect to the stack decoding at the expense of a limited increase of decoding time.
- The factored trigger model gives a limited, sometimes negligible, improvement.
- The addition of the dev and test data has little and inconsistent impact; for tst2014 it slightly tends to improve performance, vice-versa for tst2013. This behavior is probably due to small differences among the data sets; we will investigate this issue, when we will get the references.
- Our first experiment testing for interactions suggests that the discriminative model performs better under the assumption that each phrase table feature is independent from one another.

[4]During the system development many more combinations of the considered elements were tested.

task	run	tst2013		tst2014	
		BLEU	TER	BLEU	TER
MT En-Fr	pr	38.20	44.83	34.24	46.75
	cn7	38.13	44.83	34.18	46.61
	cn6	37.88	45.05	33.79	47.02
	cn5	36.27	47.48	32.07	50.02
	cn4	38.16	44.90	33.98	47.03
	cn3	38.04	44.93	34.02	46.87
	cn2	37.95	45.08	33.67	47.24
	cn1	36.73	46.44	32.49	48.81
	cn9	37.89	44.98	34.03	46.86
MT De-En	pr	25.45	55.59	20.52	63.54
	cn	25.76	55.80	20.37	63.37

Table 1: Case-sensitive BLEU and TER results for FBK's submissions to the English-French and German-English MT tasks.

The contrastive run 5, was also applied into the joint submission by the EU-Bridge project[5] partners; details about the EU-Bridge system are available in a companion paper [23].

6. German-English MT task

Our German-English systems are built ontop of the baseline system (see Section 2. Each system contains one translation model, reordering model, language model, the factored trigger model and operation sequence model; these models are then combined in a standard log-linear fashion.

The training data is composed of several publicly available corpora provided in IWSLT MT evalution task, in WMT 2014 translation task. As parallel data the following corpora were taken into account: Web Inventory of Transcribed and Translated Talks (version 2014-01) (TED) [18], German-English Europarl (version 7) (EP), Common Crawl (CC), MultiUN (UN), and the News Commentary (NC) corpus as distributed by the organizers of the Workshop of Machine Translation (WMT) 2014. We used all the available monolingual corpora provided by the WMT 2014 translation task. The target side of the parallel corpora is also used to train our LMs.

	unselected			selected		
Corpus	**Segm**	**De Words**	**En Words**	**Segm**	**De Words**	**En Words**
TED	171K	3.3M	3.46M	171K	3.3M	3.46M
CC	2.4M	56M	58M	462K	10.5M	10.7M
EP	1.9M	52M	53M	188K	3.58M	3.64M
UN	162K	5.8M	5.66M	45K	1.59M	1.52M
NC	200K	5.25M	5.0M	59K	1.4M	1.3M

Table 2: Statistics of the parallel and monolingual data exploited for training our German-English systems. For the parallel data, statistics before and after data selection are reported. Symbols "M" and "K" stand for 10^6 and 10^3, respectively.

Table 2 shows the statistics of the German-English data. The average number of words per sentence in all of the above corpora is relatively lower on German side than on the English side. This is largely due to compounding, where Noun-Verb, Noun-Noun, Adjective-Noun etc. are combined together to form a larger compound. Models trained using raw German text could lead to a high out-of-vocabulary rate on unseen texts [24]. We leverage a trainable compound splitter [25], which splits a compound based on a frequency based metric. We train one compound splitter model on TED monolingual corpus (German) which contains 3.35M running words and another on source side of TED parallel corpus (German) which contains 3.2M running words. The first splitter is `aggressive` while the second model is more `passive`. Each of the selected corpora goes through these splitter models resulting in two different systems for German-English task.

Primary: We select different amount of texts from each corpus ranging from 10% to 30%. Aggressive splitting is done on source side (German) of all training, development and test corpora. As German-English language pair shows high amount of reordering we have used hierarchical phrase reordering model as given in [26]. Each system has one TM and one RM that are built on each domain i.e. a total of 5 TMs and RMs. Linear interpolation method as described in Section 4.1.3 is used to combine just the out-of-domain models i.e CC, EP, UN and NC resulting in one background TM and RM. TED TM and RM is then filled-up with the background TM and RM with a binary provenance feature. Another model that we use is a lexically driven 5-gram operation sequence model (OSM) [27] with standard feature set. The OSM model is built on the concatenation of all 5 parallel corpora. As the factored trigger model usually results in a big phrase table, we use just the TED domain to build the model. Tree-Tagger [28] throws up a lemma and a POS-tag information for each word which makes it easy to include this information as two factors of the factored trigger model.

Contrastive: The contrastive system is configured similarly to the Primary system, except that we use the passive splitter model to split the German compounds.

Results of both systems are at par with each other on 2013 and 2014 test sets. On comparing just the BLEU scores on both test sets, we see that a passive splitter is useful for 2013 test set while an aggressive splitting is required on the 2014 test set. FTM was useful for German-English pair because an offline evaluation on the development set (tst2010) showed that the primary system with the FTM gave a jump of 0.2 BLEU points over the system where we do not use FTM.

7. English-French SLT task

The sections below describe the steps followed to perform English-French speech translation. Each of the submitted translations are drawn from machine translation systems derived from the contrastive-6 MT system (Section 5), which uses stack decoding. We briefly describe the techniques applied to normalize and preprocess the ASR outputs to make

[5] `http://www.eu-bridge.eu`

them suitable for translation. We additionally provide a brief summary of a text normalization technique relying on phonemic confusion to synthesize ASR outputs for MT training. Finally, we describe our experimental results.

7.1. Preprocessing

Prior to translating ASR outputs, we perform the following normalization steps to make them compatible with our phrase-based SMT system.

Similar to the MT track, we tokenize ASR outputs using the scripts provided by Moses. After tokenization, we recase the outputs. The recaser system is trained using the Moses scripts and a 3-gram LM. The recaser model and language models are trained on a concatenation of TED and WMT News Commentary data. Finally, we insert punctuation via monotonic machine translation, similar to the approach of [29].

7.2. Phoneme-motivated Text Normalization

A SMT system trained only on transcripts and other text data results yields a search space that is inaccessible by ASR outputs that may contain errors and text normalization issues. In an ideal scenario, we would train our spoken language translation system on a combination of text corpora and speech recognition outputs with reference translations; however, a sufficiently large amount of such speech corpora is not readily available. In order to make our machine translation system more tolerant of potential ASR errors, we use a similar phoneme-motivated text normalization approach as outlined in our previous year's submission [30] to generate additional bilingual training data from the text corpora provided in the evaluation.

We adapt the MT training data into ASR-like output to anticipate ASR errors and text normalization issues during SMT model training. We do this by leveraging several components from a target ASR system. In our experiments, we use the FBK's Kaldi English ASR system, which was used in the ASR track [31]. Similar to [32], we transform the text corpora into synthetic ASR outputs by first converting the text corpora into phonemes and then "translating" each phoneme sequence back into words that more closely match the output of our ASR system. Following the exposition described in [30], we use the Festival text-to-speech engine[6] to convert each word in our ASR system's pronunciation lexicon into phoneme sequences. The word to phoneme sequence mappings are used to generate a phrase table that translates from phoneme sequences to words. We augment the word to phoneme sequence mappings with the original pronunciation entries in the ASR lexicon. We assign uniform forward and backward phrase probabilities to each phoneme sequence to word mapping in the phrase table and omit the lexical probabilities from the model. We use the phrase table and the original ASR system's 4-gram English language model [31] as components in a Moses phrase-based SMT system.

The system is tuned on the tst2010 data set: the reference transcript is converted to phonemes using the TTS system described above. Since our goal is to convert clean transcripts into synthetic ASR output, it serves as our source text. Our reference set consists of the 1-best ASR outputs from our best Kaldi ASR system, which transcribed the audio corresponding to the tst2010 transcripts. Tuning is performed to optimize BLEU via MERT.

After tuning, we convert all of the out-of-domain text corpora, aside from Common Crawl, into ASR-like output using the trained system. Each ASR-like corpus is tokenized and recased according to the steps described above. The new damaged corpora are concatenated together and used to train an English-French phrase table and reordering model, using the same training pipeline as described in Section 2. After the phrase table and reordering models are trained, we use the fill-up technique with the models trained in the MT task (Section 5).

We additionally train a monotonic phoneme-to-phoneme phrase-based SMT system to generate additional confusable pronunciations for each of the lexical entries, using a 4-gram phoneme language model and the default Moses parameters. The training is performed in a similar manner as in [32].

7.3. Experiments

We submitted six alternative translations of the ASR outputs on tst2014. Our first set of translations (pr, cn1, cn2) use the 1-best ROVER system combination provided by the organizers. Our primary system uses all of the techniques listed above. Our first contrastive system (cn1) omits the phoneme-to-phoneme pronunciation generation. Our second contrastive system (cn2) does not include any synthetic phrase table entries. Our second set of translations (cn3-5) use the same sequence of steps as those listed above. Rather than using the ROVER ASR hypothesis, we use the ASR hypothesis corresponding to FBK's primary submission in the English ASR track. Results are shown in Table 3.

In particular, we note an increase of 1 BLEU by using the ROVER outputs instead of FBK's primary system. Additionally, we see an improvement of approximately 0.3 BLEU when using our phoneme-based text normalization techniques.

[6] http://www.cstr.ed.ac.uk/projects/festival

8. References

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech

run	BLEU	TER
pr	25.39	59.53
cn1	25.29	59.64
cn2	25.08	60.15
cn3	24.23	61.63
cn4	24.28	61.65
cn5	24.00	62.02

Table 3: Case-sensitive BLEU and TER results for FBK's tst2014 submissions to the English-French SLT task.

Republic, 2007, pp. 177–180. [Online]. Available: http://aclweb.org/anthology-new/P/P07/P07-2045.pdf

[2] 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

[3] 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, pp. 111–118.

[4] A. Axelrod, X. He, and J. Gao, "Domain Adaptation via Pseudo In-Domain Data Selection," in *Conference on Empirical Methods in Natural Language Processing*, Edinburgh, United Kingdom, 2011, pp. 355–362.

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

[6] A. Rousseau, "Xenc: An open-source tool for data selection in natural language processing," *The Prague Bulletin of Mathematical Linguistics*, vol. 100, pp. 73–82, 2013.

[7] 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.

[8] 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.

[9] G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 128–135. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0217

[10] B. Haddow, "Applying pairwise ranked optimisation to improve the interpolation of translation models," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, Georgia, USA, June 2013, pp. 342–347.

[11] 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. [Online]. Available: http://dx.doi.org/10.5167/uzh-61712

[12] R. Zhang and E. Sumita, "Boosting statistical machine translation by lemmatization and linear interpolation," 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: Association for Computational Linguistics, June 2007, pp. 181–184. [Online]. Available: http://www.aclweb.org/anthology/P07-2046

[13] P. Mathur, S. Venkatapathy, and N. Cancedda, "Fast domain adaptation of smt models without in-domain parallel data," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, August 2014, pp. 1114–1123. [Online]. Available: http://www.aclweb.org/anthology/C14-1105

[14] 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.

[15] 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.

[16] W. Kim and S. Khudanpur, "Lexical triggers and latent semantic analysis for cross-lingual language model adaptation," *ACM Transactions on Asian Language Information Processing*, vol. 3, no. 2, pp. 94–112, June 2004. [Online]. Available: http://doi.acm.org/10.1145/1034780.1034782

[17] C. Lavecchia, K. Smaïli, D. Langlois, and J.-P. Haton, "Using inter-lingual triggers for Machine translation," in *8th Annual Conference of the International Speech Communication Association - INTERSPEECH 2007*.

Antwerp, Belgium: ISCA, Aug. 2007, pp. 2829–2832. [Online]. Available: http://hal.inria.fr/inria-00155791

[18] 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. [Online]. Available: http://hltshare.fbk.eu/EAMT2012/html/Papers/59.pdf

[19] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, ser. SETQA-NLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 49–57. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622110.1622119

[20] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Harvard University, Tech. Rep. TR-10-98, 1998.

[21] 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.

[22] M. L. Buis, "Stata tip 87: Interpretation of interactions in nonlinear models," *Stata Journal*, vol. 10, no. 2, pp. 305–308(4), 2010. [Online]. Available: http://www.stata-journal.com/article.html?article=st0194

[23] M. Freitag, J. Wuebker, S. Peitz, M. Huck, A. Birch, N. Durrani, P. Koehn, M. Mediani, I. Slawik, J. Niehues, E. C. A. Waibel, N. Bertoldi, M. Cettolo, and M. Federico, "Combined spoken language translation," in *Proc. of the International Workshop on Spoken Language Translation*, Lake Tahoe, California, USA, December 2014, p. to appear.

[24] N. Ruiz, A. Bisazza, R. Cattoni, and M. Federico, "FBK's Machine Translation Systems for IWSLT 2012's TED Lectures," in *International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012, pp. 61–68.

[25] 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.

[26] 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.

[27] N. Durrani, A. Fraser, H. Schmid, H. Hoang, and P. Koehn, "Can markov models over minimal translation units help phrase-based smt?" in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 399–405. [Online]. Available: http://www.aclweb.org/anthology/P13-2071

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

[29] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling punctuation prediction as machine translation," in *International Workshop on Spoken Language Translation*, San Francisco, CA, USA, Dec. 2011, pp. 238–245. [Online]. Available: http://www.mt-archive.info/10/IWSLT-2011-Peitz.pdf

[30] A. Aue, Q. Gao, H. Hassan, X. He, G. Li, N. Ruiz, and F. Seide, "Msr-fbk iwslt 2013 slt system description," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, December 2013. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=20520

[31] B. Babaali, R. Serizel, S. J. D. Falavigna, R. Gretter, and D. Giuliani, "FBK @ IWSLT 2014 - ASR track," in *Proc. of the International Workshop on Spoken Language Translation*, Lake Tahoe, California, USA, December 2014, p. to appear.

[32] Q. F. Tan, K. Audhkhasi, P. G. Georgiou, E. Ettelaie, and S. S. Narayanan, "Automatic speech recognition system channel modeling." in *INTERSPEECH*, 2010, pp. 2442–2445.