The MIT-LL/AFRL IWSLT-2012 MT System†

Jennifer Drexler, Wade Shen, Terry Gleason

MIT/Lincoln Laboratory
Human Language Technology Group
244 Wood Street
Lexington, MA 02420, USA
{j.drexler,swade,tpg}@ll.mit.edu

Tim Anderson, Raymond Slyh, Brian Ore, Eric Hansen

Air Force Research Laboratory
Human Effectiveness Directorate
2255 H Street
Wright-Patterson AFB, OH 45433
{first.last}@wpafb.af.mil

Abstract

This paper describes the MIT-LL/AFRL statistical MT system and the improvements that were developed during the IWSLT 2012 evaluation campaign. As part of these efforts, we experimented with a number of extensions to the standard phrase-based model that improve performance on the Arabic to English and English to French TED-talk translation task. We also applied our existing ASR system to the TED-talk lecture ASR task, and combined our ASR and MT systems for the TED-talk SLT task.

We discuss the architecture of the MIT-LL/AFRL MT system, improvements over our 2011 system, and experiments we ran during the IWSLT-2012 evaluation. Specifically, we focus on 1) cross-domain translation using MAP adaptation, 2) cross-entropy filtering of MT training data, and 3) improved Arabic morphology for MT preprocessing.

1. Introduction

During the evaluation campaign for the 2012 International Workshop on Spoken Language Translation (IWSLT-2012) [1] our experimental efforts centered on 1) cross-domain translation using MAP adaptation, 2) cross-entropy filtering of machine translation (MT) training data, and 3) improved Arabic morphology for MT preprocessing.

In this paper we describe improvements over our 2011 baseline systems and methods we used to combine outputs from multiple systems. For a more in-depth description of the 2011 baseline system, refer to [3].

The remainder of this paper is structured as follows. Section 2 presents our work on the MT task, and section 3 presents our work on the automatic speech recognition (ASR) and spoken language translation (SLT) tasks. In section 2 we describe our baseline MT system, the improvements made to that system over the course of this evaluation, the experiments performed to test those improvements, and our evaluation results. In section 3 we describe our existing ASR system that was applied to both the ASR and SLT tasks, and present evaluation results for those tasks.

†This work is sponsored by the Air Force Research Laboratory under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

1.1. IWSLT-2012 Data Usage

We submitted systems for the ASR task, SLT task, and English-to-French and Arabic-to-English MT tasks. In each case, we used data supplied by the evaluation for each language pair for training and optimization. For English-to-French translation, several out-of-domain corpora were used for language model training, phrase table training, and cross-entropy filtering. For Arabic, our systems were strictly limited to the TED training supplied by the evaluation.

We employ a minimum error rate training (MERT) [20] process to optimize model parameters with a held-out development set (`dev2010`). The resulting models and optimization parameters can then be applied to test data during the decoding and rescoring phases of the translation process.

2. Machine Translation

2.1. Baseline MT System

Our baseline system implements a fairly standard SMT architecture allowing for training of a variety of word alignment types and rescoring models. It has been applied successfully to a number of different translation tasks in prior work, including prior IWSLT evaluations. The training/decoding procedure for our system is outlined in Table 1. Details of the training procedure are described in [13].

2.1.1. Phrase Table Training

When building our phrase table, we applied Kneser-Ney discounting [6] to the forward and backward translation probabilities of the phrases extracted during word alignment. In the past, we have combined multiple word alignment strategies, as described in [14]. For the experiments described here, we used only IBM model 5 (see [17] and [18]) for word alignment, to keep the statistics appropriate for discounting.

Training Process	
1.	Segment training corpus
2.	Compute GIZA++, Berkeley and Competitive Linking Alignments (CLA) for segmented data [14] [15] [16]
3.	Extract phrases for all variants of the training corpus
4.	Split word-segmented phrases into characters
5.	Combine phrase counts and normalize
6.	Train language models from the training corpus
7.	Train TrueCase models
8.	Train source language repunctuation models
Decoding/Rescoring Process	
1.	Decode input sentences use base models
2.	Add rescoring features (e.g. IBM model-1 score, etc.)
3.	Merge N-best lists (if input is ASR N-best)
4.	Rerank N-best list entries

Table 1: *Training/decoding structure*

Decoding Features
$P(\mathbf{f}\|\mathbf{e})$
$P(\mathbf{e}\|\mathbf{f})$
$LexW(\mathbf{f}\|\mathbf{e})$
$LexW(\mathbf{e}\|\mathbf{f})$
Phrase Penalty
Lexical Backoff
Word Penalty
Distortion
$\hat{P}(\mathbf{E})$ – 6-gram language model
Rescoring Features
$\hat{P}_{rescore}(\mathbf{E})$ – 7-gram LM
$\hat{P}_{class}(\mathbf{E})$ – 7-gram class-based LM
$P_{Model1}(\mathbf{F}\|\mathbf{E})$ – IBM model 1 translation probabilities

Table 2: *Independent models used in log-linear combination*

2.1.2. *Language Model Training*

During the training process we built n-gram language models (LMs) for use in decoding/rescoring, TrueCasing and repunctuation. In all cases, the MIT Language Modeling Toolkit [19] was used to create interpolated Kneser-Ney LMs. Additional class-based language models were also trained for rescoring. Some systems made use of 3- and 7-gram language models for rescoring trained on the target side of the parallel text.

2.1.3. *Optimization, Decoding, and Rescoring*

Our translation model assumes a log-linear combination of phrase translation models, language models, etc.

$$\log P(\mathbf{E}|\mathbf{F}) \propto \sum_{\forall r} \lambda_r h_r(\mathbf{E},\mathbf{F})$$

To optimize system performance we train scaling factors, λ_r, for both decoding and rescoring features so as to minimize an objective error criterion. This is done using a standard Powell-like grid search performed on a development set [20].

A full list of the independent model parameters that we used in our baseline system is shown in Table 2. All systems generated N-best lists that are then rescored and reranked using either a maximum likelihood (ML) or an minimum Bayes risk (MBR) criterion.

These model parameters are similar to those used by other phrase-based systems. For IWSLT, we also add source-target word translation pairs to the phrase table that would not have been extracted by the standard phrase extraction heuristic from IBM model 5 word alignments. These phrases have an additional lexical backoff penalty that is optimized during MERT.

The `moses` decoder [21] was used for our baseline system.

This system serves as the basis for a number of the contrastive systems submitted during this year's evaluation. As described in the following sections, we implemented several techniques for generating improved phrase tables and language models, and experimented with using these techniques both individually and in combination.

2.2. English-To-French Domain Adaptation

During this evaluation we re-examined the approach to cross domain adaptation that we presented in last year's evaluation [3]. Instead of training a single out-of-domain model to adapt to the TED domain, we trained individual models for each available parallel corpus and combined them using hierarchical MAP adaptation [2]. In this technique, models trained on corpora that are more distant from the test domain are successively MAP-adapted with models estimated from less distant corpora, using the following equation:

$$\begin{aligned}\hat{p_i}(s|t,\lambda) = &\frac{N_i(s,t)}{N_i(s,t)+\tau_i}p_i(s|t,\lambda_i)\\ &+\frac{\tau_i}{N_i(s,t)+\tau_i}\hat{p_{i+1}}(s|t,\lambda_{i+1}) \quad (1)\end{aligned}$$

where $N_i(s,t)$ is the count of the phrase pair (s,t) in model i, $p_i(s|t,\lambda_i)$ is the probability of the source phrase given the target phrase in model i, and $\hat{p_{i+1}}(s|t,\lambda_{i+1})$ is the MAP estimate from the previous step. The final probability estimate for the given phrase pair is $\hat{p}_1(s|t)$. The full hierarchy can be seen in Figure 1.

For the experiments presented here, the ordering of the MAP hierarchy was determined based on the BLEU score of each individual translation model on the held-out TED development set, with low-scoring models adapted towards higher-scoring ones.

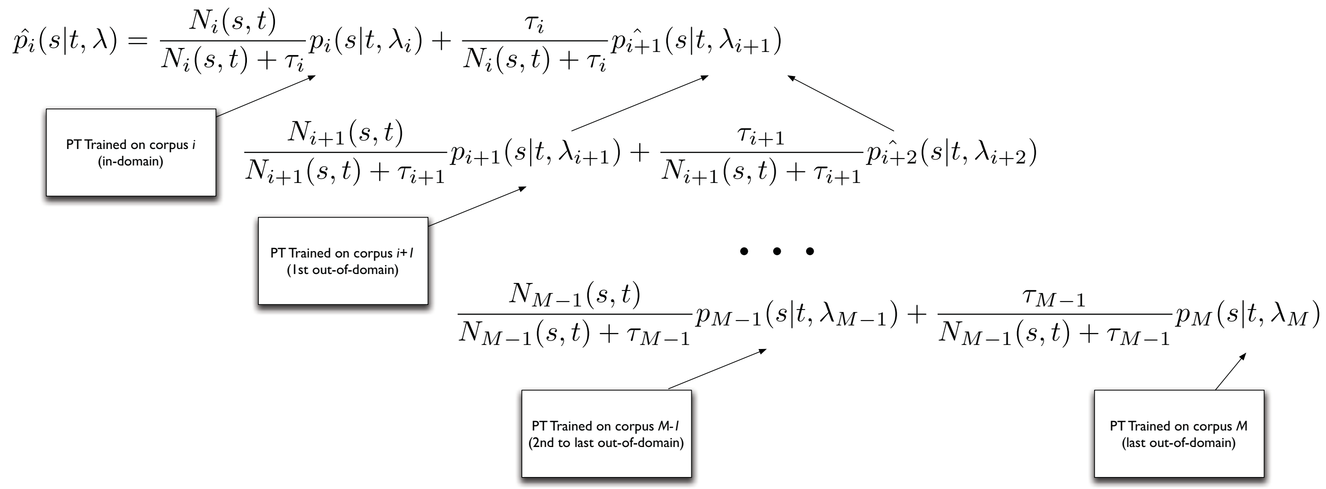

Figure 1: MAP with multiple corpora

2.3. English-To-French Cross-Entropy Filtering

As a comparison to domain adaptation, we experimented with cross-entropy training data filtering, as in [38]. We tested both language model- and translation model-based filtering, but used only LM-based filtering for the experiments performed here, as we found no significant improvement from the inclusion of translation model scores.

We performed LM cross-entropy filtering separately on the parallel portions of the Europarl, Giga-FrEn, News Commentary, and UN corpora. For each of these corpora, for both the source and target sides, we trained a language model on a random subset of the sentences of the same size as the TED training data. We then sorted all sentences in the corpus based on the difference between their cross-entropy given this model and their cross-entropy given the TED language model. We trained new language models on the best $1/64$, $1/32$, $1/16$, $1/8$, $1/4$, and $1/2$ of the corpus. We selected the filter size that produced the language model with the minimum perplexity on the `dev2010` dataset.

To filter the parallel data, we combined the perplexity thresholds that produced the best source and target language models for the `dev2010` dataset. This resulted in the selection of 3.2 percent of the overall data for translation model and language model training, as shown in Table 3.

Two translation models were trained using the filtered parallel data. For the first, which we refer to as A3part, the alignments were generated using all the filtered data but then only the alignments from the TED portion were used to build the translation model. For the second, called TMFilt, the translation model was fully generated from all of the filtered data.

2.4. Alternate French Language Models for Rescoring

Continuous space language model (CSLM) [37], and recurrent neural network language model (RNNLM) [36] were trained on the target side of the TED data. The continuous space language model contained 256 hidden units and an input context of 4 words. The recurrent neural network contained 160 hidden units, 300 classes and backpropagation through time of 4. These language models were used as additional rescoring models on the n-best list. A recurrent neural network language model was also trained on the target side of the bilingual cross-entropy filtered data (RNN-TMfilt). Another language model used for rescoring was the maximum entropy language model(MELM). The 3-gram language model was adapted from a background MELM trained on gigaword and TED data. These models were trained with an extension of the SRILM toolkit.

Corpus	Before Filtering	After Filtering
TED	141,387	141,387
Giga-FrEn	24,116,560	824,698
UN	12,886,831	220,066
Europarl	2,007,723	76,554
News Commentary	137,097	1,735
TOTAL	39,289,598	1,264,441

Table 3: *Cross-entropy filtering results in term of number of sentence pairs*

2.5. Arabic Morphological Processing

In our Arabic-to-English MT systems for prior year evaluations [10, 9, 8, 7, 3], we normalized various forms of alef and hamza and removed the tatweel character and some diacritics before applying a light Arabic morphological analysis procedure that we called AP5. This year, we modified the AP5 procedure to more closely conform to the Arabic Treebank (ATB) segmentation format used in the MADA Arabic morphological analysis, diacritization, and lemmatization system

		Arabic	English
`train`	Sentences	90,542	
	Running words	1,235,359	1,477,768
	Avg. Sent. length	13.64	16.32
	Vocabulary	46,780	34,447
`dev2010`	Sentences	934	
	Running words	13,719	17,451
	Avg. Sent. length	14.68	18.68
`tst2010`	Sentences	507	
	Running words	23,080	26,786
	Avg. Sent. length	13.87	16.10
		English	**French**
`train`	Sentences	141,387	
	Running words	2,356,136	2,468,430
	Avg. Sent. length	16.66	17.46
	Vocabulary	41,466	53,997
`dev2010`	Sentences	934	
	Running words	17,451	17043
	Avg. Sent. length	18.68	18.25
`tst2010`	Sentences	1664	
	Running words	26,786	27,802
	Avg. Sent. length	16.10	16.71

Table 4: *Corpus statistics for all language pairs*

[4]. In [5], it was shown that the ATB format performed the best of the various MADA segmentation formats tried on the IWSLT 2011 evaluation. In particular, we kept the definite article (Al-) attached to its corresponding noun or adjective. We denote this modified AP5 system as AP5ATBLite.

2.6. MT Experiments

With each of the enhancements presented in prior sections, we ran a number of development experiments in preparation for this year's evaluation. This section describes the development data that was used for each evaluation track, and results comparing the aforementioned enhancements with our baseline system.

2.6.1. Development Data

Tables 4 describes the development and training set configurations used for each language pair in this year's evaluation. We used the WMT-supplied segmenters for preprocessing and normalization, as well as in-house tokenizers for Arabic and French.

2.6.2. English-to-French MT Experiments

We ran a number of baseline and experimental systems on the talk task data set using the methods described in prior sections. In order to perform development experiments, we used supplied development data (`dev2010`) for optimization, and we held out `tst2010` for development testing. Table 5 summarizes the results on the held-out `tst2010` set. For these experiments, the reported scores are an average of ten optimization/decoding runs with different random weight initializations. In all cases we use at at least a 6-gram LM for decoding and rescore with a 7-gram class LM and model1.

Table 5 contains results of our experiments with training data filtering, and with the use of additional language models for rescoring. The three sections of this table show results obtained with three different phrase tables. The first of these, the baseline phrase table, was generated using only the supplied TED training data. The next phrase table, A3Part, was generated using the cross-entropy filtering method described in Section 2.3. Specifically, the word alignments were generated using all of the filtered data, but the phrases were extracted only from the TED data. This phrase table gives an improvement of more than one BLEU point over the baseline. The last phrase table, referred to as TMFilt, was again generated from the filtered data, this time using all of the data for both word alignment and phrase extraction. This phrase table gives an additional improvement of more than half a BLEU point over the A3part phrase table.

Within each section of Table 5, the experiments differ based on their language model configurations. The baseline TED language model was used in all cases. For all except the first line in each section, a language model trained from the monolingual Gigaword data was also used. This language model is a 6-gram language model interpolated by year over the afp portion of the French Gigaword corpus. It adds more than half a BLEU point, regardless of the phrase table it is used with. We also show results using additional language models (CSLM, RNN, MELM) for rescoring. These language models provided little or no additional gain in performance, and in one case reduced the overall gain.

System	`tst2010`
TED Models Only (baseline)	32.06
TED PT + InterpGiga LM	32.61
A3part	33.16
A3part + InterpGiga LM	33.80
A3part + InterpGiga LM + RNN	33.57
A3part + InterpGiga LM + MELM	33.79
A3part + InterpGiga LM + CSLM	33.91
A3part + InterpGiga LM + CSLM + RNN-TMfilt	33.83
TMFilt	33.71
TMFilt + InterpGiga LM	34.22
TMFilt + InterpGiga LM + RNN	34.26
TMFilt + InterpGiga LM + MELM	34.35
TMFilt + InterpGiga LM + CSLM	34.40
TMFilt + InterpGiga LM + CSLM + RNN-TMfilt	34.24

Table 5: *Summary of English-French filtering experiment results*

Table 6 contains results from our domain adaptation experiments. The MAP phrase table was produced through hierarchical MAP adaptation of phrase tables trained with the following parallel corpora (in order): News Commentary, Europarl, Giga-FrEn, and TED. On its own, this phrase table improves the baseline score by about half a BLEU point. We combined our phrase table domain adaptation with language models that were trained individually on each parallel corpus and included in the log-linear model. Using these language models adds an additional half BLEU point to our scores.

System	`tst2010`
TED Models Only (baseline)	32.06
TED PT + Parallel LMs	32.58
MAP	32.60
MAP + Parallel LMs	33.27

Table 6: *Summary of English-French domain adaptation experiment results*

The overall best result was achieved with the TMFilt phrase table, when combined with rescoring using a CSLM language model. This score, 34.40, represents a gain of 2.34 BLEU points over the baseline score of 32.06. Unfortunately, the TMFilt phrase table results were generated too late to be included in the evaluation. At submission time, our best individual system used the same configuration, but with the A3Part phrase table instead of the TMFilt phrase table, for an average BLEU score of 33.91.

As described in section 2.7, we were able to combine our domain adaptation system with one of our filtering systems to produce a better result than any of the individual systems available at submission time. In the future, we plan to experiment with ways of combining the best techniques from domain adaptation and filtering into a single system, rather than relying on system combination.

2.6.3. *Arabic-To-English MT Experiments*

Table 7 shows the mean BLEU scores for individual Arabic-to-English MT systems trained on the 2011 and 2012 training data and tested on the `tst2010` data versus the morphology segmentation system. For both the 2011 and 2012 training data, the AP5ATBLite system performs slightly better than the AP5 system. Also, the extra training data in the 2012 system provides approximately one BLEU point of improvement over the systems trained on the 2011 data.

Table 7: *Mean BLEU scores for individual Arabic-to-English MT systems tested on the tst2010 data versus morphology segmentation system and year of training data.*

Morphology System	Training Data	
	2011	2012
AP5	21.13	22.24
AP5ATBLite	21.57	22.45

In addition to the AP5ATBLite modification, we investigated the use of Kneser-Ney (KN) phrase table smoothing [6] using the AP5ATBLite system trained on the 2012 training data. The combination of AP5ATBLite and KN smoothing yielded a mean BLEU score of 23.60 compared to the mean of 22.45 for the AP5ATBLite system without phrase table smoothing.

2.7. MT Submission Summary

As part of this year's evaluation we experimented with training data filtering, improved cross-domain adaptation, and improved Arabic morphological processing. These developments have helped to improve our system when compared with our 2011 system.

The overall submitted Arabic-to-English system was a combination of individual component systems that were each the best in terms of BLEU score after ten MERT optimization runs. Two of the component systems were (1) the best AP5ATBLite system (with no phrase table smoothing) and (2) the best AP5ATBLite system with KN phrase table smoothing.

The majority of our English-To-French submissions are also combinations of multiple systems. Our primary submission is a combination of the *MAP + Parallel LMs* system and the *A3part + InterpGiga LM + MELM* system. We also submitted the individual system that had the best single MERT run, in terms of BLEU score on the `tst2010` data set, which was a run of the *A3part + InterpGiga LM + CSLM + RNN-TMfilt* system.

Table 8 summarizes each of the systems submitted for this year's evaluation and how they compare with our 2011 submission (when applicable) on the `tst2011` and `tst2012` data sets. Due to a de-tokenization error, our official English-to-French submissions had much lower scores; the scores reported here reflect the performance of our system after the correction of that error.

3. Automatic Speech Recognition and Spoken Language Translation

3.1. ASR System

Acoustic models were developed using the same TED data and training procedure as our IWSLT 2011 system [3]. In addition to training models using Perceptual Linear Prediction (PLP) features, we trained a second set of acoustic models using Mel-Frequency Cepstral Coefficients (MFCCs).

Cross-entropy difference scoring [35] was used to select subsets of the Europarl, Gigaword, news 2007–2011, and news commentary texts for training the language models. The provided TED training data was used for the in-domain text, and the selection threshold for each out-of-domain data set was chosen to minimize the perplexity on `dev2010`. This process selected 7.3% of the data for LM development.

The SRILM Toolkit[1] was used to estimate interpolated

[1]Available at: http://www.speech.sri.com/projects/srilm

Arabic-to-English Systems			
System	*Features*	`tst2011`	`tst2012`
AE-primary 2011	2011 combined system	19.56	N/A
AE-primary	2012 primary combination	17.99	19.30
AE-contrast1	2012 contrast1	17.28	18.36
English-to-French Systems			
System	*Features*	`tst2011`	`tst2012`
EF-primary 2011	2011 best system	34.19	N/A
EF-primary	2012 primary combination	36.10	37.32
EF-contrast1	2012 best individual system	36.16	36.75
EF-contrast2	2012 best combination	36.39	37.10

Table 8: *Summary of submitted 2012 MT systems*

trigram and 4-gram LMs for decoding and rescoring, respectively. Recurrent Neural Network Maximum Entropy (RNNME) LMs [36] were developed for rescoring using the RNNLM Toolkit.[2] One RNNME LM was trained on Gigaword, and a second RNNME LM was trained on news 2007–2011. As suggested in [39], the number of classes was set to 300 and 4-gram features were used for the ME model. Each network included 160 hidden units, which was selected to minimize the perplexity on the TED training data.[3] The vocabulary for the LMs included 95,000 words.

Recognition lattices were produced using the same procedure as last year [3], and 1000-best lists were extracted for rescoring with the 4-gram and RNNME LMs. The scores from each LM were linearly interpolated using weights chosen to minimize the perplexity on the development partitions. The final transcripts were produced by combining the MFCC and PLP systems using a Confusion Network Combination system (CNC).[4]

Our implementation of CNC starts by creating confusion networks for each recognizer's rescored N-best list. These confusion networks are then aligned to each other using a time-weighted Levenschtein distance computed over the max posterior hypothesis per recognizer. The resulting alignment is used to merge columns of each individual confusion network into a single confusion network, where language model and acoustic model scores for each recognizer's hypotheses are combined in a log-linear way, with weights for each system and each individual model. System weights were set through a Powell-like grid search using the supplied development data.

Table 9 shows the Word Error Rates (WERs) obtained on the IWSLT `dev2010` and `tst2010`partitions. According to the unofficial results, the submitted system yielded a 12.6% WER on `tst2011` and a 14.3% WER on `tst2012`.

[2] Available at: http://www.fit.vutbr.cz/∼imikolov/rnnlm

[3] Due to time constraints we only compared networks with 80, 120, and 160 hidden units.

[4] Due to a bug in the submitted system, the submitted combination did not result in significant differences between the PLP baseline and the submitted combination. This was due to an error in setting the prior weight per system.

	`dev2010`		`tst2010`	
	MFCC	PLP	MFCC	PLP
1st pass	19.0	18.3	18.7	17.9
2nd pass	16.6	16.5	15.4	15.0
4-gram	15.3	15.4	14.1	13.9
4-gram + RNNME	14.4	14.4	13.0	12.5
CN combination	13.7		12.9	

Table 9: WERs obtained on the IWSLT `dev2010` and `tst2010` partitions using the MFCC and PLP systems.

3.2. SLT System

For the SLT task, we used a combination of the ASR and MT systems described above. We used only ASR input from our own system.

3.3. SLT Submission

Table 10 summarizes the results of our submission for the SLT tasks. Our official SLT evaluation scores were impacted by the same de-tokenization error that lowered our English-to-French MT scores. Again, these scores reflect the performance of our system once that error was corrected.

System	`tst2011`	`tst2012`
Primary	27.82	27.54
Contrastive	27.52	27.51

Table 10: *Summary of submitted 2012 SLT systems*

4. Acknowledgments

We would also like to thank Katherine Young and Jeremy Gwinnup for their help in processing the English-French and TED task data sets and the staff of the Human Language Technology group at MIT Lincoln Lab for making machines available for this evaluation effort.

5. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, S. Stüker "Overview of the IWSLT 2012 Evaluation Campaign," In *Proc. of IWSLT*, Hong Kong, HK, 2012.

[2] A. R. Aminzadeh, J. Drexler, T. Anderson, and W. Shen, "Improved Phrase Translation Modeling Using MAP Adaptation," in *Proceedings of TSD 2012* (Brno, Czech Republic), September 2012.

[3] A. R. Aminzadeh, T. Anderson, R. Slyh, B. Ore, E. Hansen, W. Shen, J. Drexler, and T. Gleason, "The MIT-LL/AFRL IWSLT-2011 MT system," in *Proceedings of IWSLT 2011,* (San Francisco CA), December 2011.

[4] R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin, "Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking," in *Proceedings of ACL-08: HLT, Short Papers,* (Columbus OH), June 2008.

[5] J. Wuebker, M. Huck, S. Mansour, M. Freitag, M. Feng, S. Peitz, C. Schmidt, and H. Ney, "The RWTH Aachen machine translation system for IWSLT 2011," in *Proceedings of IWSLT 2011,* (San Francisco CA), December 2011.

[6] G. Foster, R. Kuhn, and H. Johnson, "Phrasetable smoothing for statistical machine translation," in *Proceedings of EMNLP 2006,* (Sydney, Australia), July 2006.

[7] Shen, Anderson, T., Slyh, R., and Aminzadeh, A.R., "The MIT-LL/AFRL IWSLT-2010 MT System," In Proc. Of the International Workshop on Spoken Language Translation, Paris, France, 2010.

[8] Shen, W., Delaney, B., Aminzadeh, A.R., Anderson, T., and Slyh, R. "The MIT-LL/AFRL IWSLT-2009 MT System," In Proc. Of the International Workshop on Spoken Language Translation, Tokyo, Japan, 2009.

[9] Shen, W., Delaney, B., Anderson, T., and Slyh, R. "The MIT-LL/AFRL IWSLT-2008 MT System," In Proc. Of the International Workshop on Spoken Language Translation, Honolulu, HI, 2008.

[10] Shen, W., Delaney, B., Anderson, T., and Slyh, R. "The MIT-LL/AFRL IWSLT-2007 MT System," In Proc. Of the International Workshop on Spoken Language Translation, Trento, Italy, 2007.

[11] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," In Proc. of MT Summit, 2005.

[12] Munteanu, D. S. and Marcu, D., "ISI Arabic-English Automatically Extracted Parallel Text," Linguistic Data Consortium, Philadelphia, 2007.

[13] Shen, W., Delaney, B., and Anderson, T. "The MIT-LL/AFRL IWSLT-2006 MT System," In Proc. Of the International Workshop on Spoken Language Translation, Kyoto, Japan, 2006.

[14] Chen, B. et al, "The ITC-irst SMT System for IWSLT-2005," In Proc. Of the International Workshop on Spoken Language Translation, Pittsburgh, PA, 2005.

[15] Melamed, D., "Models of Translational Equivalence among Words," In Computational Linguistics, vol. 26, no. 2, pp. 221-249, 2000.

[16] Liang, P., Scar, B., and Klein, D., "Alignment by Agreement," Proceedings of Human Language Technology and North American Association for Computational Linguistics (HLT/NAACL), 2006.

[17] Brown, P., Della Pietra, V., Della Pietra, S. and Mercer, R. "The Mathematics of Statistical Machine Translation: Parameter Estimation," Computational Linguistics 19(2):263–311, 1993.

[18] Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, I.D., Och, F.J., Purdy, D., Smith, N.A., Yarowsky, D., "Statistical machine translation: Final report," In Proceedings of the Summer Workshop on Language Engineering at JHU, Baltimore, MD 1999.

[19] Bo-June (Paul) Hsu and James Glass, "Iterative Language Model Estimation: Efficient Data Structure and Algorithms," In Proc. Interspeech, 2008.

[20] Och, F. J., "Minimum Error Rate Training for Statistical Machine Translation," In ACL 2003: Proc. of the Association for Computational Linguistics, Japan, Sapporo, 2003.

[21] Koehn, P., et al, "Moses: Open Source Toolkit for Statistical Machine Translation," Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic, June 2007.

[22] K. Oflazer and I. Kuruoz, "Tagging and morphological disambiguation of Turkish text," In Proceedings of the 4th Conference on Applied Natural Language Processing, Stuttgart, Germany, 1994.

[23] Mermer, C., Kaya, H., and Dogan, M.U. "The TUBITAK-UEKAE Statistical Machine Translation System for IWSLT 2007," In Proc. of IWSLT, 2007.

[24] Matusov, E. and Ueffing, N. and Ney, H., "Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment," In Proc. of EACL, 2006.

[25] Fiscus, JG, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 1997.

[26] Snover, M. and Dorr, B. and Schwartz, R. and Micciulla, L. and Makhoul, J., "A study of translation edit rate with targeted human annotation," In Proc. of AMTA, 2006.

[27] Rosti, A.V.I. and Matsoukas, S. and Schwartz, R., "Improved Word-Level System Combination for Machine Translation," In Proc. of ACL, 2006.

[28] T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki "Online large-margin training for statistical machine translation," In Proc. of EMNLP-CoNLL, 2007.

[29] D. Chiang Y. Marton, and P. Resnik, "Online large-margin training of syntactic and structural translation features," In Proc of EMNLP, 2008.

[30] D. Chiang, K. Knight, W. Wang, "11,001 new features for statistical machine translation," In Proc. NAACL/HLT, 2009.

[31] D. Graff, J. Garofolo, J. Fiscus, W. Fisher, and D. Pallett, "1996 English Broadcast News Speech (HUB4)," *Lingustic Data Consortium*, Philadelphia, 1997. Available: http://www.ldc.upenn.edu

[32] J. Fiscus, J. Garofolo, J. Fiscus, M. Przybocki, W. Fisher, and D. Pallett, "1997 English Broadcast News Speech (HUB4)," *Lingustic Data Consortium*, Philadelphia, 1998. Available: http://www.ldc.upenn.edu

[33] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[34] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008.

[35] R. Moore and W. Lewis, "Intelligent Selection of Language Model Training Data," *Association Computational Linguists 2010 Conference Short Papers*, Uppsala, Sweden.

[36] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černocký, "Strategies for Training Large Scale Neural Network Language Models," in *Proc. Automatic Speech Recognition and Understanding Workshop*, Hawaii, USA, 2011.

[37] Schwenk, Holger, "Continuous Space Language Models," in *Computer Speech and Language*, vol 21, 492-518, 2007.

[38] S. Mansour *et al.*, "Combining Translation and Language Model Scoring for Domain-Specific Data Filtering," in *Proc. International Workshop on Spoken Language Translation*, San Francisco, USA, 2011.

[39] Mikolov Tom, Karafit Martin, Burget Luk, ernock Jan, Khudanpur Sanjeev, "Recurrent neural network based language model," In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, JP, 2010.