The MIT-LL/AFRL IWSLT-2011 MT System

A. Ryan Aminzadeh

Department of Defense
ryan.aminzadeh@ugov.gov

Tim Anderson, Ray Slyh, Brian Ore, Eric Hansen

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

Wade Shen, Jennifer Drexler, Terry Gleason[†]

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

Abstract

This paper describes the MIT-LL/AFRL statistical MT system and the improvements that were developed during the IWSLT 2011 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 tasks. We also applied our existing ASR system to the TED-talk lecture ASR task.

We discuss the architecture of the MIT-LL/AFRL MT system, improvements over our 2010 system, and experiments we ran during the IWSLT-2011 evaluation. Specifically, we focus on 1) speech recognition for lecture-like data, 2) cross-domain translation using MAP adaptation, and 3) improved Arabic morphology for MT preprocessing.

1. Introduction

During the evaluation campaign for the 2011 International Workshop on Spoken Language Translation (IWSLT-2011) our experimental efforts centered on 1) speech recognition for lecture-like data, 2) improved cross-domain translation using MAP adaptation using corpus distance measures in addition to count-based smoothing, and 3) improved Arabic morphology for MT preprocessing.

In this paper we describe improvements over our 2010 baseline systems and methods we used to combine outputs from multiple systems. For a more full description of the 2010 baseline system, refer to [1].

The remainder of this paper is structured as follows. In section 2, we present an overview of our baseline system and the minor improvements to this standard statistical MT architecture that we developed. In sections 3, 4, and 5 we describe experiments for cross-domain adaptation, better Arabic morphological processing. Section 7 describes the systems we submitted for this year's evaluation and their results.

1.1. IWSLT-2011 Data Usage

We submitted systems for the ASR 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 systems, data from Gigaword and Europarl corpora were used for both language model and phrase model training. For Arabic, our systems were strictly limited to the TED training supplied by the evaluation.

In order to train models based on the Gigaword corpus, we randomly sampled 15% of the total corpus for training phrase models. In order to minimize compute time, we eliminated sentences longer than 40 words. The entire data set was used to train language models.

For cross-domain adaptation experiments conducted on the English-to-French data sets, the TED training data was used to adapt these initial models to the TED domain(s).

We employ a minimum error rate training 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. 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 [7].

2.1. Phrase Table Training

To maximize phrase table coverage, we combine multiple word alignment strategies, extending the method described in [8]. For all language pairs, we combine alignments from IBM model 5 (see [11] and [12]) with alignments extracted using the competitive linking algorithm (CLA) described

[†]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.

	Training Process
1.	Segment training corpus
2.	Compute GIZA++, Berkeley and Competitive Linking Alignments (CLA) for segmented data [8] [9] [10]
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*

in [9] and the Berkeley Aligner [10]. Phrases were extracted from both types of alignments and combined in one phrase table. This was done by summing counts of phrases extracted from alignment types before computing the relative frequencies used in our phrase tables.

2.2. Language Model Training

During the training process we built n-gram language models for use in decoding/rescoring, TrueCasing and repunctuation. In all cases, the MIT Language Modeling Toolkit [13] was used to create interpolated Knesser-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.

2.3. 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 [14].

In addition to the Powell-based approach, a number of our systems used the MIRA algorithm for weight optimization [23, 22, 24]. In this approach, weights are optimized subject to a maximum margin constraint in an online fashion. The equation below shows the update procedure for weights w_i corresponding to the ith online iteration of the algorithm.

$$\mathbf{w_i} = \mathbf{w_{i-1}} + \alpha * (\mathbf{h}(f, \hat{e}) - \mathbf{h}(f, e))$$

where $\hat{e}$ denotes the oracle translation for a source sentence f, $\mathbf{h}(f, e)$ is a vector of model scores corresponding to the translation of f into e, and α is an update scaling parameter defined as follows:

$$\alpha = max(0, min(C, \frac{\mathcal{L}(\hat{e}, e) - (s^{i-1}(f, \hat{e}) - s^{i-1}(f, e))}{||\mathbf{h}(f, \hat{e}) - \mathbf{h}(f, e)||}))$$

$$s^{i-1}(f, e) = \mathbf{w_{i-1}} \cdot \mathbf{h}(f, e)$$

$\mathcal{L}(\hat{e}, e)$ defines a loss function (in our case, the BLEU score difference between the oracle translation, $\hat{e}$, and the current best translation, e. C is a limiter on the update scaling. It's easy to see that update size at each iteration is proportional to the difference between the loss value and the predicted score margin.

Weights $\mathbf{w_i}$ are updated sentence by sentence (order of presentation is randomized) until either a convergence criterion is met or a limit on the number of iterations is reached. Our implementation of MIRA follows the procedure in [23] for oracle selection and scoring.

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 ML (Maximum Likelihood) or an MBR (Minimum Bayes Risk) criterion.

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*

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 minimum error rate training.

This system serves as the basis for a number of the contrastive systems submitted during this year's evaluation. Contrastive systems differ in terms of their rescoring configuration (e.g. language models, MBR) and the data used to train them (some systems made use of additional lexicon data). Each of the contrastive systems was used as a component for system combination. The combined output for each of the English-to-French and Arabic-to-English tasks was submitted as our primary system. Detailed differences of each submitted system can be found in section 8.

The `moses` decoder [15] was used for our baseline system.

3. Automatic Speech Recognition

Acoustic training data for our ASR system were harvested from TED. We downloaded 807 TED Talks that were recorded prior to 2011, and used FFmpeg to extract 16 kHz audio from each video file. Word alignments were automatically generated for each talk using an HTK HMM system that was trained on the HUB4 English Broadcast News corpora [25, 26]. Long periods of non-speech were removed, and each talk was split into utterances shorter than 20 seconds. Next, closed caption filtering [27] was applied to sequester utterances that may include transcription errors. Each talk was decoded using the HUB4 HMMs and a Language Model (LM) that was estimated from the transcript for the talk. The recognizer outputs were compared to the transcripts, and a data partition was created using all utterances with a Word Error Rate (WER) less than 20%. This process yielded 164 hours of audio.

An HMM system was trained on the TED acoustic data using HTK. Phonemes were modeled using state-clustered cross-word triphones, and the final HMM set included 6,000 shared states with an average of 28 mixtures per state. The models were discriminatively trained using the Minimum Phone Error (MPE) criterion. The feature set consisted of 12 Perceptual Linear Prediction (PLP) coefficients, plus the zeroth coefficient, with mean normalization applied on a per utterance basis. Delta, acceleration, and third differential coefficients were appended to form a 52 dimensional vector, and Heteroscedastic Linear Discriminant Analysis (HLDA) was applied to reduce the feature dimension to 39. A second set of models was estimated that included Speaker Adaptive Training (SAT).

Interpolated LMs were trained from the Europarl, News commentary, News 2007–2011, and the provided TED data. Trigram and 4-gram LMs were estimated for decoding and rescoring. The vocabulary included 95,000 words, and unknown pronunciations were added to the CMU dictionary using the Sequitur grapheme-to-phoneme system [28].

Decoding was performed as follows. Initial transcripts were produced using HDecode with the non-SAT HMMs. Next, the MIT-LL GMM software package was used to cluster the utterances from each talk. Constrained Maximum Likelihood Linear Regression (CMLLR) transforms were estimated for each cluster, and recognition lattices were generated using the SAT HMMs. The final transcripts were produced by rescoring the lattices with the 4-gram LM.

4. Cross Domain Adaptation

During this evaluation we re-examined the approach to cross domain adaptation that we presented in last year's evaluation [2]. To this end, we built a general purpose model for the English-French task using training data from the Gigaword French-English and Europarl corpora [5] for each language respectively. These models were trained using over 700k sentence pairs of data. Using the provided training data from the IWSLT evaluation, we applied a variation of the MAP phrase table adaptation procedure described last year, which is shown in the equations below:

$$\begin{aligned}\hat{p}(s|t) &= \lambda p_{iwslt}(s|t) + (1-\lambda)p_{gp}(s|t)\\ \lambda &= \frac{N_{iwslt}(s,t)}{N_{iwslt}(s,t)+N_{gp}(s,t)+\tau} * \hat{p_0}(iwslt|dev)\end{aligned}$$

where p_{gp} and p_{iwslt} are phrase probability estimates from the general purpose and IWSLT-domain models respectively, and $p_0(iwslt|dev)$ is an estimate of the corpus posterior given a dev set.

During last year's evaluation we used a strict MAP formulation, i.e., the ratio of counts between $iwslt$ and gp models determines the weighting of the models. During this evaluation we introduce a corpus posterior probability p_0 which we approximate via dev set BLEU scores as follows:

$$\hat{p_0}(dev|iwslt) \approx \frac{BLEU(dev|\lambda_{iwslt})}{\forall_c BLEU(dev|\lambda_c)}$$

where $BLEU(dev|\lambda_c)$ is simply the BLEU scores for a dev set dev under a translation model λ_c. The idea here is to incorporate a semantic distance measure to weight the contribution of phrase counts from each corpus.

As in last year's experiments, phrase table adaptation and language model interpolation were used jointly to improve performance.

5. Arabic-specific Morphological Processing

In our Arabic systems for prior year evaluations [4, 3, 2, 1], we normalized various forms of alef and hamza and removed the tatweel character and some diacritics before applying a light Arabic morphological analysis procedure. This year, we first normalized all Unicode Arabic presentation forms to their constituent isolated forms. For example, Unicode `\x{fef7}` (called "Arabic Ligature Lam with Alef with Hamza Above Isolated Form") was normalized to Unicode `\x{0644}\x{0623}` (*i.e.,* "Arabic Letter Lam" in isolated form followed by the isolated form for "Arabic Letter Alef with Hamza Above"). Then, we performed our alef-hamza, tatweel, and diacritic normalizations. At this time, we further removed Unicode `\x{0670}`, "Arabic Letter Superscript Alef," and normalized Unicode `\x{0671}`, "Arabic Letter Alef Wasla," to Unicode `\x{0627}`, a "bare" alef. After these normalizations, we converted Arabic digits and the Arabic percent sign, decimal separator, and thousands

		dev2010	tst2010
HUB4 HMMs	1st pass	25.3	24.8
	2nd pass	23.1	21.2
	4-gram rescore	22.6	20.7
TED HMMs	1st pass	20.7	19.7
	2nd pass	18.5	16.4
	4-gram rescore	17.8	15.8

Table 3: WERs obtained on the IWSLT dev2010 and tst2010 partitions using the HUB4 and TED HMMs

separator to their English equivalents and tokenized the punctuation.

In our 2009 Arabic MT system [2], we employed a modification of our earlier light morphological analysis process that we called Count-Mediated Morphological Analysis (CoMMA). The CoMMA process segments only those tokens that occur in the training data fewer times than a user-chosen threshold. Tokens that occur at least as many times as the threshold are passed through to the output unsegmented. For this year's Arabic system, we again employed the CoMMA process and developed six MT systems using the CoMMA process at thresholds of 1,000, 2,000, 5,000, 10,000, 20,000, and 50,000. For each of these six threshold values, the best system in terms of BLEU score (after ten optimization runs) was used in our system combination with the other Arabic MT systems that we developed.

6. ASR Experiments

Table 3 shows the WERs obtained on the IWSLT dev2010 and tst2010 partitions. For comparison purposes, we have also included the results obtained with the HUB4 HMMs. Note that non-SAT HMMs were used for both passes with the HUB4 system. From Table 3, we can see that training on the TED acoustic data yielded a substantial improvement in WER compared to the HUB4 models.

7. 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.

7.1. Development Data

Tables 4 describes the development and training set configurations used for each language pair in this year's evaluation.

7.2. English-to-French Translation Baselines

We ran a number of baseline systems on the talk task data set using the methods described in prior sections. We used the WMT-supplied segmenters for preprocessing and normalization, as well as in-house tokenizers for Arabic and French. In addition to the IWSLT-supplied TED data, data from the French Gigaword and Europarl corpora was used for language/phrase modeling in the English-French task (our Arabic-English system makes no use of non-TED data). 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 five optimization/decoding runs with different random weight initializations.

		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	107,268	
	Running words	1,760,288	1,840,764
	Avg. Sent. length	16.41	17.16
	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*

No single optimization strategy clearly outperforms the other, though the addition of language models trained on other corpora is a clear benefit (≈1.0-1.2 BLEU). During this evaluation we employed a perplexity-minimizing interpolation strategy: a single LM was constructed by interpolating TED LMs with LMs trained on other corpora so as to minimize perplexity on dev2010.

7.2.1. English-French Domain Adaptation Experiments

As described in section 4, we applied a different formulation of the MAP-based count-smoothing approach we introduced during last year's evaluation and, in this year's evaluation, we also introduced a corpus-distance factor. We conducted experiments on the English-to-French translation task using out-of-domain data from Europarl and Gigaword corpora for

System	Optimization Method	`tst2010`
TED PT + TED LM	MERT	29.54
TED PT + TED LM	MIRA	29.12
TED PT + TED LM + additional LMs	MERT	30.80
TED PT + TED LM + additional LMs	MIRA	31.07

Table 5: *Summary of baseline TED English-French translation task experiments*

backoff when in-domain model probabilities are poorly estimated.

Table 6 compares the English-to-French IWSLT baseline (optimized via a Powell search) against, 1) the MAP adaptation method we proposed last year and 2) MAP with a corpus distance factor as proposed above.

In both cases, a gain of ≈1 BLEU point can be had. Intuitively, by using relative counts, the new approach allows more refined computation of the λ used to compute the interpolated/adapted probability for each phrase. This method avoids overweighting the gp model when both the $iwslt$ and gp models have relatively few counts.

For these experiments, the reported scores are an average of five optimization/decoding runs with different random weight initializations. Note that both variants of our phrase table adaptation result in gains over language model interpolation alone. The use of a corpus-based distance measure in addition to the standard MAP approach results in a small ≈0.2-0.4 BLEU gain on the supplied `tst2010` data set, but the results on the `tst2011` data set don't show significant differences. This could be due to mismatch between `dev2010` (which was used to compute corpus weights for interpolation) and `tst2011`. More experiments will be needed to explore this performance gap.

7.3. Arabic Morphology Experiments

We evaluated the translation results from the modified CoMMA processes, as described above, for the Arabic-to-English translation task at the aforementioned threshold levels. Table 7 shows the mean BLEU scores (over ten optimization runs) on the the Arabic `tst2010` development data set when applying CoMMA. These systems were then compared to a baseline using the unmodified CoMMA procedure as described in last year's system.

The revised IWSLT11 CoMMA process did not consistently outperformed the standard CoMMA process in a BLEU signifcant manner, though at most threshold points there was a slight gain. We would have expected that the revised normalization should allow for more consistent Arabic phrase extraction, but this didn't results in large BLEU score gains, perhaps due to the relatively large training set available training.

CoMMA	Mean BLEU	
Threshold	CoMMA (Old)	CoMMA (New)
1,000	20.40	20.27
2,000	20.18	20.26
5,000	20.33	20.44
10,000	20.23	20.44
20,000	21.06	22.18
50,000	21.52	22.10

Table 7: *Mean BLEU scores for IWSLT10 and IWSLT11 CoMMA systems versus threshold for the Arabic* `tst2010`

8. Evaluation Summary

As part of this year's evaluation we experimented with improved cross-domain adaptation, improved Arabic morphological processing and refinements to our multiple MT combination approach. These developments have helped to improve our system when compared with our 2010 baseline.

Table 8 summarizes each of the systems submitted for this year's evaluation and how they compare with our 2010 baselines (when applicable) on the `tst2011` data set.

9. 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.

10. References

[1] 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.

[2] 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.

[3] 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.

System	`tst2010`	`tst2011`
TED Model Only (baseline)	29.54	28.53
TED Model Only + Additional LMs (baseline)	30.80	31.12
TED MAP-adapted ([1])	31.91	33.78
TED MAP-adapted (BLEU-based corpus-distance)	32.23	33.81

Table 6: *Summary of adaptation experiment results*

Arabic-to-English Systems		
System	*Features*	BLEU
AE-primary	2011 combined system	19.56
AE-contrast2	2011 best individual system (CoMMA t=50,000)	19.47
English-to-French Systems		
System	*Features*	BLEU
EF-primary 2010	2010 baseline	31.12
EF-primary	2011 combined system	34.19
EF-contrast2	2011 best individual system (domain-adapted PT + LM)	33.81

Table 8: *Summary of submitted 2011 systems*

[4] 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.

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

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

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

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

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

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

[25] 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

[26] 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

[27] 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.

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