The University of Washington Machine Translation System for the IWSLT 2007 Competition

Katrin Kirchhoff and Mei Yang

Department of Electrical Engineering
University of Washington
{katrin,yangmei}@ee.washington.edu

Abstract

This paper presents the University of Washington's submission to the 2007 IWSLT benchmark evaluation. The UW system participated in two data tracks, Italian-to-English and Arabic-to-English. Our main focus was on incorporating out-of-domain data, which contributed to improvements for both language pairs in both the clean text and ASR output conditions. In addition, we compared supervised and semi-supervised preprocessing schemes for the Arabic-to-English task and found that the semi-supervised scheme performs competitively with the supervised algorithm while using a fraction of the run-time.

1. Introduction

We describe University of Washington's machine translation system for the 2007 IWSLT competition. Our system participated in two tracks, the Italian-to-English Challenge task and the Arabic-to-English Classical task. For Italian-to-English translation our main focus was on utilizing out-of-corpus training resources and to determine the relative benefit of having small amounts of in-domain training data vs. larger amounts of unrelated data. For Arabic-to-English translation, we also investigated using text resources unrelated to the BTEC travel task. In addition, we compared linguistic methods for tokenization vs. a semi-supervised algorithm that seeks to improve initial tokenization by utilizing unannotated data. The following sections describe the basic setup of the translation system (Section 2), the Italian-to-English System (Section 3), the Arabic-to-English system (Section 4), and experiments and results (Section 5).

2. Translation System

2.1. Translation model

Our system is a phrase-based statistical MT system based on a log-linear probability model:

$$e^* = argmax_e p(e|f) = argmax_e \{\sum_{k=1}^{K} \lambda_k \phi_k(e,f)\} \quad (1)$$

Given and English sentence e and a foreign sentence f, $\phi(e,f)$ is a feature function defined on both sentences, and λ is a feature weight. In particular, we use the following feature functions:

- two phrase-based translation scores, one for each translation direction
- two lexical translation scores, one for each translation direction
- word count penalty
- phrase count penalty
- distortion penalty
- language model score
- a data source feature

Phrasal and lexical translation scores are computed as shown below in Equations 2 and 3. For a segmentation of source and target sentences into phrases, $f = \bar{f}_1, \bar{f}_2, ..., \bar{f}_M$ and $e = \bar{e}_1, \bar{e}_2, ..., \bar{e}_M$, the phrasal translation score for $\bar{e}$ given $\bar{f}$ is computed as

$$P(\bar{e}|\bar{f}) = \frac{count(\bar{e}, \bar{f})}{count(\bar{f})} \quad (2)$$

i.e. as the relative frequency estimate from the phrase-segmented training corpus. The lexical score is computed as

$$Score_{lex}(\bar{e}|\bar{f}) = \prod_{j=1}^{J} \frac{1}{|\{j|a(i)=j\}|} \sum_{a(i)=j}^{I} p(f_j|e_i) \quad (3)$$

where j ranges over words in phrase $\bar{f}$ and i ranges over words in phrase $\bar{e}$.

Phrases are extracted from the word-aligned training corpus using the heuristic technique described in [6]. For word alignment we use an HMM-based word to phrase alignment model [1]. Under this model, target phrases are generated by individual source words, including the NULL word. The alignment of source words and target phrases is governed by a first-order Markov process. For a target sentence f of length m, segmented into K phrases, a source sentence e of length l, and a sequence of alignment variables a_1^K, the alignment model is specified as

$$P(a_1^K, h_1^K, \phi_1^K | K, m, e) = \prod_{k=1}^{K} P(a_k, h_k, \phi_k | a_{k-1}\phi_{k-1}, e) \prod_{k=1}^{K} p(a_k | a_{k-1}, h_k; l) d(h_k) n(\phi_k; e_{a_k})$$

where h_1^K is a series of binary variables indicating whether a phrase is inserted or not, ϕ_1^K are variables controlling the length of the target phrase, $n(\phi; e)$ is a simple length model for a word e producing a phrase of length ϕ, and d is an i.i.d. process with $d(0) = p_0, d(1) = 1 - p_0$. We use the word-to-phrase alignment as implemented by the MTTK package [2]. A previous comparison against GIZA++ based word alignment on the IWSLT 2006 training corpus for Italian-to-English found a marginal improvement of this model over GIZA++.

Word count and phrase count penalties are constant weights added for each word/phrase used in the translation; the distortion penalty is a weight that increases in proportion to the number of positions by which phrases are reordered during translation. The language model score is obtained from a trigram trained using SRILM [11]. The weights for these scores are optimized using an in-house implementation of the minimum-error rate training procedure developed in [7]. Our optimization criterion is the BLEU score on the development set.

For both systems we use additional out-of-corpus data sources. As described in [3], these are integrated into the system by training separate phrase tables on each data source and using both tables jointly during decoding without renormalization. An additional feature function (or functions, in the case of more than two data sources) in the log-linear model indicates which data source a phrase pair was extracted from; the weight for this feature is optimized along with all other feature weights to maximize the BLEU score on the development set. Identical phrase pairs extracted from different data sources are included in the phrase table multiple times, along with their different scores and respective data source features. We found this method to be helpful on previous IWSLT tasks.

2.2. Decoding

For decoding we use the Moses package [4] in its basic form, i.e. without making use of any of its advanced features such as factored translation models. We utilize two decoding passes, a first pass that generates up to 2000 hypotheses per sentence, and a second pass that rescores the initial hypotheses by utilizing additional model scores. For both systems we use a part-of-speech based trigram language model in the second pass. The parts-of-speech were annotated using a maximum-entropy tagger for English [8]. Additional models are specific to each language pair and are discussed in the following sections. Model weights are reoptimized for the second decoding pass. For both language pairs, the number of possible positions by which phrases may be reordered is limited to four.

	BTEC	Europarl
# sentence pairs	26,467	625,320
average # words	160K	17M

Table 1: Sizes of the Italian datasets.

2.3. Postprocessing

The output from the second decoding pass is postprocessed to restore true case and punctuation. We use a hidden-event n-gram model [9, 10] to restore punctuation and a noisy-channel model for truecasing. The hidden-ngram model implements a statistical language model over an event set E consisting of regular words and an additional set of punctuation signs.

$$P(e_1, ..., e_T) \approx \prod_{t=n}^{T} P(e_t | e_{t-1}, ..., e_{t-n+1}) \qquad (4)$$

During training, all events are observed; during testing, hidden events are hypothesized after every word. Their posterior probability is computed by using a forward-backward dynamic programming procedure and the transition probabilities provided by n-gram model trained on punctuated text. The noisy-channel model is a 4-gram model trained over a mixed-case representation of the BTEC training corpus and a probabilistic mapping table for lowercase-uppercase word variants. It was implemented using the *disambig* tool from the SRILM package [11].

3. Italian-to-English System

3.1. Data

For our Italian-to-English translation model we use two data sources: the BTEC training corpus and the Europarl corpus of parliamentary proceedings. We use all available BTEC data for training (i.e. including the development sets for previous IWSLT competitions but not the 2007 development set). Data sizes are shown in Table 1. The Europarl data is of a fundamentally different nature than the BTEC corpus since it consists of transcriptions of parliamentary proceedings. Its style is that of written text and it is much larger than the BTEC corpus. For development we used the 2007 development set, which was randomly split into an actual development set of 500 sentences and a held-out set of 496 sentences.

3.2. Preprocessing

The BTEC data was split by segmenting lines with multiple sentences into smaller chunks based on punctuation marks.

Subsequently, punctuation signs were removed and all data was lowercased. Due to the small training data size, word alignments were observed to be noisy. We sought to improve the word alignment by automatic re-tokenization of both the Italian and the English text. For the top twenty words that are aligned to multiple adjacent words we merged the multiply aligned words into a single item prior to training the word aligner. This increased the BLEU score on the held-out set by 0.5% for a baseline of 18.4%.

3.3. Out-of-vocabulary words

We observed OOV words in the development set that seem to be indicative of the spoken nature of the task, e.g. *senz'*, *undic'*, *quant'*, etc. We use a general procedure to map OOV words to their closest counterparts in the training data by means of string alignment. Words are compared to all words in the training vocabulary that differ in length by not more than 2 characters. Each of these words is then aligned to the unknown word by means of dynamic programming and their edit distance is computed. For all candidates with an edit distance less than 2, the corresponding phrase table entries are extracted and reduplicated with the word in question replaced by the unknown word. Thus, during decoding, the best-matching phrase table entry according to the translation and the language model can be selected. This technique often effectively finds the correct forms for misspelled words as well as spoken-language specific elisions. Although the overall impact on the translation score can be expected to be low (due to the low percentage of OOVs) we expect the translation of OOVS to positively influence human evaluation.

3.4. Data Combination

For data combination, the phrase table trained on the Europarl corpus was used as-is, without retraining the word aligner on the pooled BTEC and Europarl data. This was done because the small size of the BTEC corpus makes it unlikely to influence the quality of the word alignment and subsequent phrase extraction to a significant degree.

3.5. Rescoring features

For rescoring we use a rank-based feature in addition to a POS-based language model. The rank feature (see [3]) indicates the rank of the hypothesis in the nbest list output by the first decoding pass, and also ties together identical hypotheses generated by different phrase segmentations. The value of this feature is equivalent to the position of the hypothesis in the N-best list unless an identical, higher-ranked hypothesis has already been found. In that case, it takes on the value of the higher-ranked hypothesis. This feature was found to be beneficial in our IWSLT 2006 experiments and was re-used in this year's system.

	BTEC	News texts
# English words	153,053	1.2M
# Arabic words	161,207	5.5M
# sentence pairs	23,176	190,140

Table 2: Sizes of data sets used for the AE task.

3.6. Translation of names and numbers

We used the BTEC list of proper names provided on the IWSLT 2007 resources web page in order to annotate named entities. These were kept fixed and were not translated by the phrase table. Numbers such as dates and times were translated by a small number of translation rules rather than statistically.

3.7. Spoken-language specific processing

No specific spoken-language processing was performed in our system. Initial experiments with using confusion network input instead of the 1-best ASR hypothesis did not show any significant gains, so that the 1-best ASR hypothesis was used in all cases. The feature function weights in the log-linear translation model were optimized on the clean text and were not re-optimized for the ASR condition.

4. Arabic-to-English System

4.1. Data

For our Arabic-to-English system we used the provided BTEC training corpus with the exception of the dev4 and dev5 sets, which were used for system development. We additionally used the allowable parallel Arabic text corpora from LDC (Arabic Newswire, Multiple-Translation Arabic, and automatically extracted parallel text provided by ISI), which consist of news texts and their translations. The corpus sizes are shown in Table 2 and refer to the number of tokens after preprocessing.

Additionally, we use the knowledge base of the Buckwalter stemmer available from the LDC in that we use its list of stems and their English translations as part of the training data.

4.2. Preprocessing

The training data was chunked according to punctuation signs and converted to Buckwalter transliteration. We then compared two different schemes for tokenization: one based on rule-based linguistic analysis and one semi-supervised algorithm. Linguistic preprocessing is done using the Buckwalter stemmer and the Columbia University MADA and TOKAN tools. The Buckwalter analyzer proposes a number of different morphological analyses of a word form, based on a list of word stems and morphological rules. The MADA tool then statistically chooses one particular analysis based

Tool	BLEU/PER
supervised	22.5/50.5
semi-supervised	23.0/50.7

Table 3: First pass BLEU(%)/PER scores for the AE dev5 set, without true case/punctuation (clean text).

on the surrounding context. The TOKAN tool uses this analysis to split off clitics and particles such as *w-, b-, etc.* In particular, we used the setting "*w+ f+ l+ b+ k+ Al+* REST", i.e. all word initial particles as well as the definite article were split off.

The development of tokenization tools such as the ones described above require significant human labour and expertise. We are interested in evaluating the performance of automatic or semi-automatic tokenization algorithms against linguistic baselines, which may be useful for porting MT systems to new languages or dialects. We therefore used the semi-supervised approach to Arabic tokenization presented in [12], which was initially developed for dialectal Arabic, for which standard analyzers do not exist. This algorithm starts from (a) a small seed set of words segmented into prefixes, stems, and suffixes, and (b) a list of affixes. Words not in the seed set are then attempted to be segmented by removing possible affixes. Resulting ambiguous segmentations are resolved by applying stem frequency information (the segmentation with the more frequent stem is chosen). This procedure is applied iteratively, each time updating the seed set with new stems and segmented words. It is thus a way of extending and improving the initial segmentation hypotheses automatically using a larger set of unannotated data. This is similar to the method presented in [5]; however, it is even less data-intensive since no statistical language model is used. We were interested in applying this technique to the IWSLT data to assess its performance on unseen data, in particular unseen words that may not be analyzable by the Buckwalter tool. The semi-supervised algorithm was initialized with the linguistic segmentations on the BTEC training set and the trained segmenter was used to re-segment the training, development and held-out sets.

Two systems trained on the linguistic vs. semi-supervised segmentations were compared. Their performance was roughly the same, as shown in Table 3. However, once trained, the semi-supervised segmenter runs in about 30% of the time required by the linguistic annotation tools since it mostly works with look-up tables.

4.3. Rescoring

For rescoring we use a 4-gram POS-based language model only. The rank feature used in the Italian-to-English system did give an improvement on the AE development set but did not generalize to the held-out set and was not used in the final system.

	Clean Text						
Corpus	1	2	3	4	5	6	7
BTEC	78.2	29.6	6.7	1.3	0.2	0	0
Europarl	83.9	37.0	6.4	0.7	0.1	0	0
combined	85.9	39.9	9.4	1.7	0.2	0	0

Table 4: Phrase coverage (in %) of the IE 2007 development set (clean text) for different data sources.

	ASR Output						
Corpus	1	2	3	4	5	6	7
BTEC	76.8	26.6	5.4	0.9	0.1	0	0
Europarl	86.5	36.8	6.2	0.6	0.1	0	0
combined	91.7	47.7	10.9	1.5	0.2	0	0

Table 5: Phrase coverage (in %) of the IE 2007 development set (ASR output) for different data sources.

5. Experiments and Results

5.1. Italian-to-English

We first investigated the coverage of phrases of different lengths obtained by the individual vs. the combined data sources. Tables 4 and 5 show the percentages of phrases out of the 2007 development set covered under the different conditions. In all cases, coverage of phrases longer than two words is extremely poor. At the same time, the coverage of 1-word and 2-word phrases is increased significantly by combining multiple data sources. This holds for both the clean text and the ASR condition. Coverage improvement is high in the ASR condition; however, this does not necessarily impact MT performance since the additional words covered by adding the Europarl data may be recognition errors.

In order to further determine the usefulness of the out-of-domain (OOD) data compared to in-domain data we trained a phrase table on the 500 non-held-out sentence of the 2007 development set. Note that this system did not contribute to the results submitted for the challenge task evaluation, whose goal it was to evaluate *cross-domain* performance. Rather, it serves as a point of comparison in order to judge how systems perform when training data consists of

- a small amount of in-domain data, vs.
- a moderate amount of domain-related but stylistically different data, vs.
- a large amount of data different in domain and style, vs.
- a weighted combination of data sources.

We thus compared the performance of the system on the held-out part of the development set when trained (a) only on the in-domain data set, (b) only on the read-speech BTEC

	Data Source(s)	heldout set text
a	in-domain BTEC set only	28.0/46.8
b	out-of-domain BTEC set only	18.9/55.1
c	Europarl only	18.5/55.4
d	out-of-domain BTEC + Europarl	20.7/53.5
e	all combined	30.1/41.9

Table 6: First-pass BLEU(%)/PER scores for systems trained on different data sources, without true case/punctuation (clean text).

	Data Source(s)	heldout set ASR
a	in-domain BTEC set only	26.1/49.0
b	out-of-domain BTEC set only	16.6/57.1
c	Europarl only	17.3/56.4
d	out-of-domain BTEC + Europarl	18.6/55.3
e	all combined	27.7/44.8

Table 7: First-pass BLEU (%)/PER scores for systems trained on different data sources, without true case/punctuation (ASR output).

training data (c) only on Europarl, (d) on both read-speech BTEC and Europarl, and (e) on all corpora combined.

Tables 6 and 7 show that the in-domain trained system clearly is the best individual system, even though its training set is very small. Interestingly, the system trained on a moderately sized corpus of read speech from the travel domain only performs marginally better than a system trained on the unrelated Europarl corpus. The combination of all data sources further improves over the best individual system by a significant amount. The improvement holds up in the ASR condition – this is different from previous results on the IWSLT 2006 IE task, where out-of-domain data contributed to better system performance for clean text but not for ASR output. The reason is most likely the better performance of the ASR front-end for the IWSLT 2007 task.

Table 8 shows the results of the various system development steps on the held-out set.

Step	BLEU/PER
Baseline	18.9/55.1
Addition of OOD data	20.7/53.4
Rescoring	22.0/52.7
Dates/Numbers	22.6/52.2
True case & punctuation	21.2/50.4

Table 8: BLEU (%)/PER for various system development steps (IE system, clean text).

	Clean Text						
Corpus	1	2	3	4	5	6	7
BTEC	69.3	34.3	12.9	3.6	1.2	0.3	0
news text	60.2	30.4	11.5	2.7	0.9	0.2	0
combined	82.6	46.7	20.1	5.6	1.9	0.5	0.1

Table 9: Phrase coverage (in %) of the AE dev5 set (clean text) for different data sources.

	ASR output						
Corpus	1	2	3	4	5	6	7
BTEC	66.0	33.3	12.3	3.5	1.1	0.3	0
news text	69.3	12.5	0.8	0	0	0	0
combined	75.9	32.7	9.5	2.2	0.6	0.1	0

Table 10: Phrase coverage (in %) of the AE dev5 set (ASR output) for different data sources.

5.2. Arabic-to-English

As before, we measured the improvement in coverage when adding out-of-domain data. Tables 9 and 10 show the results. Again, we see a significant improvement obtained from the combined training set.

The use of out-of-domain data yielded the largest improvements, whereas rescoring hardly improved the performance.

Table 11 shows the results of the various system development steps on the held-out set.

Finally, Table 12 shows the results obtained in the official evaluation.

6. Conclusions

We have presented the UW MT system for the IWSLT 2007 competition. Our main focus was on integrating out-of-domain data (news texts and parliamentary proceedings). The additional data helped improve the system performance for both languages and in both translation conditions (clean text and ASR output). For the IE Challenge task we showed that a large amount of general purpose text covering a wide range of topic can achieve approximately the same performance as a moderately sized corpus of domain-related train-

Step	text	ASR output
BTEC training only	22.5/50.5	19.1/53.8
BTEC + text data	24.6/48.2	21.0/52.0
Rescoring	24.6/47.5	21.1/51.3
True case & punctuation	23.4/48.5	20.3/52.8

Table 11: BLEU(%)/PER scores on the AE dev5 set for various system development steps (clean text).

	Eval set BLEU (%) scores
IE - Clean text	26.51
IE - ASR output	25.40
AE - Clean text	41.62
AE - ASR output	40.92

Table 12: Final evaluation results.

ing data.

7. Acknowledgments

This work was funded by NSF Grant IIS-0308297 from the U.S. National Science Foundation.

8. References

[1] Y. Deng and W. Byrne, "HMM Word and phrase alignment for statistical machine translation", *Proceedings of HLT-EMNLP*, 2005

[2] W. Byrne and Y. Deng, "MTTK: an alignment toolkit for statistical machine translation", *Proceedings of HLT-NAACL*, 2006, pp. 265-268

[3] K. Kirchhoff et al., "The University of Washington Machine Translation System for IWSLT 2006", *Proceedings of IWSLT 2006*

[4] P. Koehn et al., "Moses: open source toolkit for statistical machine translation", *Proceedings of ACL*, 2007

[5] Y.S. Lee, K. Papineni and S. Roukos, "Language Model Based Arabic Word Segmentation", *Proceedings of ACL*, 2003, pp. 399-406

[6] Och, F.J., and Ney, H., "A systematic comparison of various statistical alignment models", Computational Linguistics 29(1), 19-52, 2003

[7] Och, F.J., "Minimum Error Rate Training for Statistical Machine Translation", in Proc. of 41st Meeting of the Association for Computational Linguistics, 2003.

[8] Ratnaparkhi, A., "A maximum entropy part-of-speech tagger", in Proc. of Empirical Methods in Natural Language Processing (EMNLP), 1996.

[9] Stolcke, A. and Shriberg, E., "Statistical language modeling for speech disfluencies", Proc. of Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1996, 405-409

[10] Stolcke, A. and Shriberg, E., "Automatic linguistic segmentation of conversational speech", Proc. of Int'l Conf. on Spoken Language Processing (ICSLP), 1996, 1005-1008

[11] Stolcke, A., "SRILM - an extensible language modeling toolkit", Proc. of Int'l Conf. on Spoken Language Processing (ICSLP), 2002,901-904

[12] M. Yang, A. Kathol and K. Precoda, "A semi-supervised learning approach for morpheme segmentation for an Arabic dialect", *Proceedings of Interspeech*, 2007