The University of Edinburgh System Description for IWSLT 2007

Josh Schroeder, Philipp Koehn

School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW
Scotland, United Kingdom
j.schroeder@ed.ac.uk, pkoehn@inf.ed.ac.uk

Abstract

We present the University of Edinburgh's submission for the IWSLT 2007 shared task. Our efforts focused on adapting our statistical machine translation system to the open data conditions for the Italian-English task of the evaluation campaign. We examine the challenges of building a system with a limited set of in-domain development data (SITAL), a small training corpus in a related but distinct domain (BTEC), and a large out of domain corpus (Europarl). We concentrated on the corrected text track, and present additional results of our experiments using the open-source Moses MT system with speech input.

1. Introduction

The IWSLT 2007 shared task offered the University of Edinburgh a chance to expand our experience with spoken language translation and with translation using data from multiple corpora. We focused on the Italian-English challenge task because it offered a chance to explore spontaneous speech as well as an opportunity to use a corpus we are familiar with, Europarl[1], as an additional data resource.

In this paper we first present a summary of the phrase-based statistical machine translation system used for this shared task. We go on to discuss the data sources we used to train the system, and show the results of our analysis of the domains of the development test data sets compared to the corpora used for training.

Having explained the framework and data used, in Section 5 we present the results of our experiments in cross-domain adaptation. In Section 6 we describe the experiments we conducted with ASR inputs to our system.

2. Framework

2.1. The Moses MT system

The open source Moses[2] MT system was originally developed at the University of Edinburgh. It received a major boost through a 2006 Johns Hopkins workshop, and is now used at several academic institutions as the basic infrastructure for statistical machine translation research.

The Moses system is an implementation of the phrase-based machine translation approach[3]. In this approach, an input sentence is first split into text chunks (so-called phrases), which are then mapped one-to-one to target phrases using a large phrase translation table. Phrases may be reordered, but typically a maximum movement reordering limit is used.

Phrase translation probabilities, reordering probabilities and language model probabilities are combined to give each possible sentence translation a score. The best-scoring translation is searched for by the decoding algorithm and outputted by the system as the best translation. The different system components h_i (phrase translation probabilities, language model, etc.) are combined in a log-linear model to obtain the score for the translation **e** of an input sentence **f**:

$$score(\mathbf{e}, \mathbf{f}) = \exp \sum_i \lambda_i \, h_i(\mathbf{e}, \mathbf{f}) \quad (1)$$

The weights of the components λ_i are set by minimum error rate training on held-out development data[4]. The basic components used in our experiments are:

- two phrase translation probabilities (both $p(e|f)$ and $p(f|e)$)
- two word translation probabilities (both $p(e|f)$ and $p(f|e)$)
- phrase count
- output word count
- language model
- distance-based reordering model
- lexicalised reordering model

For a more detailed description of this model, please refer to [5].

2.2. Lexicalised reordering

There are various models for reordering words to match the target language's word order. A simplistic method is distance based reordering, which uses a factor δ^n to penalise movements over n words. Moses also implements a more sophisticated method called lexicalised reordering[5] which pays

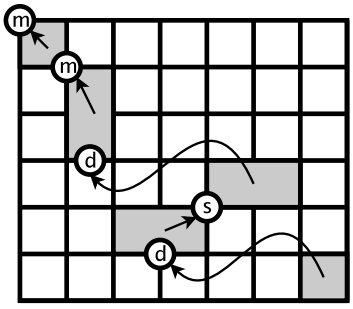

Figure 1: *Possible orientations of phrases: monotone (m), swap (s), or discontinuous (d).*

attention to the identity of the phrases being used. In lexicalised reordering, we learn for each phrase pair how likely it is to directly follow the previous phrase (monotone), to swap positions with the previous phrase (swap), or to not be connected at all with the previous phrase (discontinuous). An illustration of this is provided in Fig. 1.

We use bidirectional reordering, taking both the next and previous translated phrase into account [6]. As described above, phrase pairs are collected and classified based on their reordering type relative to other phrase alignments generated from the sentence pair:

- monotone: a word alignment point to the top left exists
- swap: an alignment point to the top right exists
- discontinuous: no alignment points to the top left nor top right

We can use the counts from this classification to learn a smoothed orientation probability distribution:

$$p_r(\text{orientation}|\bar{e}, \bar{f}) \quad (2)$$

3. Training data

The IWSLT evaluation campaign for 2007 featured only the open data condition. One set of training data was provided in the BTEC domain. Five development sets were provided in the BTEC domain and one in the SITAL domain. For this evaluation, we chose to focus on the cleaned text transcriptions of the ASR data.

3.1. Corpora

A small BTEC corpus was provided. We used the raw Europarl data available online[1] to align over 800,000 Italian-English sentences and construct a corpus providing additional coverage for our translation system. See Table 1 for full corpora statistics.

3.2. Development data

Six development sets were provided for tuning and testing. Three of the sets (devsets 1, 2 and 3) did not provide ASR inputs. We focused on devsets 4, 5a, and 5b which contained non-punctuated Italian inputs and cased, punctuated English outputs. The inputs consisted of both ASR and cleaned text data. We converted the SLF inputs to a confusion network format that could be read by the Moses decoder. Additionally, we used the 1-best text input as another form of ASR data.

[1] http://www.statmt.org/europarl

Table 1: *Statistics for BTEC and Europarl training corpora and extracted phrase tables.*

BTEC	**Italian**	**English**
Sentences	19,972	
Words	147,564	188,961
Phrase table entries	314,874	
EUROPARL	**Italian**	**English**
Sentences	868,047	
Words	22,586,316	25,267,363
Phrase table entries	49,018,026	

3.3. Punctuation adjustments

There is no punctuation in the BTEC and SITAL development sets that were derived from ASR. We addressed this by stripping punctuation from the source side for both Europarl and BTEC training corpora, thus creating an MT system that translates from un-punctuated Italian to punctuated English. The target side language model was capable of judging most punctuation re-insertion issues, and we added a small post-processing script to eliminate multiple punctuation and ensure final punctuation.

4. Phrase table coverage and domain perplexity

A condition unique to the IWSLT 2007 Challenge Tasks is that the test set is not from the same domain as the training corpus. The BTEC corpus and development sets are tourism-related sentences, and the SITAL data consists of dialogues between customers and phone operators at a travel agency. To analyse how similar the SITAL data was to the BTEC and Europarl corpora, we compared test sets from each of the 3 domains:

- Europarl: WMT07 test2007 (2000 lines)[2]
- BTEC: IWSLT07 devset4 (489 lines) and devset5a (500 lines)
- SITAL: IWSLT07 devset5b (996 lines)

[2] For the Europarl test set, we extracted the Italian sentences that were aligned with the English test set sentences used in the ACL 2007 Workshop on Machine Translation shared task. The shared task at this workshop used only English, Spanish, French, and German Europarl data, so we used the heldout portion of the Europarl corpus to locate the Italian versions of the same test set sentences.

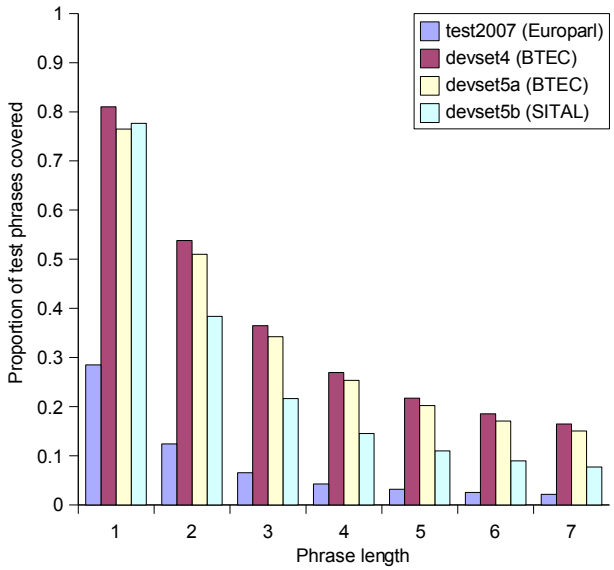

Figure 2: *Phrase coverage within the BTEC phrase table for each test set (test set domain in parentheses).*

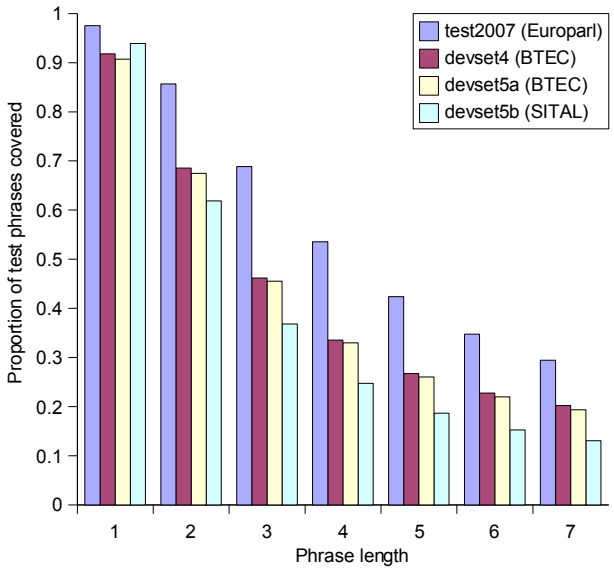

Figure 3: *Phrase coverage within the Europarl phrase table for each test set (test set domain in parentheses).*

By using in-domain test sets as a baseline, we can see how much coverage each corpus's phrase table gives to the SITAL data. One way to measure how close our test data is to our training data is to look at the perplexity with respect to a language model. We trained a 5-gram language model using SRILM[7] on the source side of each of the BTEC and Europarl corpora, and computed the perplexity of each of the test sets. This is shown in Table 2.

Table 2: *Source side language model perplexity: test2007 is out of domain for the BTEC corpus, devset4 and devset5a are out of domain for Europarl. Devset5b is out of domain for both corpora.*

Test Set	Test Set Domain	LM Corpus	Perplexity
test2007	Europarl	BTEC	982.876
devset4	BTEC	BTEC	171.67
devset5a	BTEC	BTEC	184.161
devset5b	SITAL	BTEC	311.835
test2007	Europarl	Europarl	94.2004
devset4	BTEC	Europarl	1294.4
devset5a	BTEC	Europarl	1139.26
devset5b	SITAL	Europarl	1868.88

From these two measures we can see that language model perplexity is much lower for the in-domain conditions. This gives us an idea of the "distance" between the domains of each test set. Devsets 4 and 5a are both in the BTEC corpus domain, and they have a similarly low perplexity with respect to that language model. As expected, the test2007 set has the lowest perplexity on the Europarl language model when compared to other sets. Interestingly, devset5b does not have as low of a perplexity as devsets 4 and 5a on either LM. This would seem to indicate that there is a difference in language between the SITAL data and the BTEC data.

In Figs. 2 and 3, we examine this relationship from the perspective of the phrase tables. We enumerated all phrases up to length 7 (our phrase table maximum) in each test set, and recorded what percentages of the unique phrases in each set were covered by the BTEC and Europarl phrase tables. Again, we see that the overall best coverage for a test set is provided by its in-domain phrase table (Europarl for test2007 and BTEC for devset4 and devset5a). In general, devset5b has fewer matches than the in-domain sets for both phrase tables.

A subtle difference between devset5b and the BTEC test data (devset4, devset5a) is revealed in the unigram coverage for each test set. Devset5b actually has better single-word coverage than the two BTEC sets when measured on the Europarl phrase table, and falls between them on the BTEC phrase table. However, the coverage drops lower than the BTEC sets as soon as we consider phrases of two or more words. While the single word vocabulary has better coverage, multi-word phrases are not matched as well.

We can see a possible explanation by looking at the number of unique words in each of the test sets. To control for set length, we examine the unigram and bigram counts only for the first 4,976 words of each file (the shortest test set, devset4, is 4,976 words long). This is shown in Table 3.

The SITAL set, devset5b, has less than half the number of unique words compared to the vocabulary of the other test sets. While many of these single words are present in both the BTEC and Europarl phrase tables, the bigrams they form are not. This may be a symptom of the devset5b SITAL data being spontaneous speech. Devset5b has a smaller set of vo-

Table 3: *Unique unigram and bigram phrase table coverage in the first 4,976 words of each test set.*

UNIGRAM COVERAGE			
Test Set	**Unique Unigrams**	**BTEC Coverage**	**Europarl Coverage**
test2007	1737	788 (45.4%)	1721 (99.1%)
devset4	1234	1000 (81.0%)	1133 (91.8%)
devset5a	1331	1040 (78.1%)	1212 (91.1%)
devset5b	600	497 (82.8%)	564 (94.0%)
BIGRAM COVERAGE			
Test Set	**Unique Bigrams**	**BTEC Coverage**	**Europarl Coverage**
test2007	5747	1361 (23.7%)	5226 (90.9%)
devset4	4537	2441 (53.8%)	3110 (68.5%)
devset5a	4789	2498 (52.2%)	3303 (69.0%)
devset5b	2984	1292 (43.3%)	1900 (63.7%)

cabulary, but those words are arranged in n-grams that are uncommon in both the BTEC and Europarl styles of language.

5. Experiments in cross-domain adaptation

Given that the official test set for the evaluation would be in a domain not ideally covered by either corpus, we attempted to make all the data from each separate corpus available to the decoder. This can be looked at as a form of mixture modelling or cross-domain adaptation. There has been extensive research in the area of adapting SMT systems to new domains. Recent work has distinguished between cross-domain adaptation, where the domain of the test data is known ahead of time, and dynamic adaptation, where the system must adapt to the test domain on-the-fly without the ability to tune ahead of time[8]. In both cases, models from each domain-specific corpus are trained separately, and weighted relative to their fit with the test domain.

In this case, the domain of the test set (SITAL) is known ahead of time. Given that there was only one set of development test data in the domain of the final test set, we chose to split devset5b into two equal halves, using one half for tuning and one half for testing. We refer to these sets below as devset5b-tune and devset5b-test. To minimise vocabulary shifts between individual dialogue sessions and speakers, we shuffled the sentences before splitting.

5.1. Single corpus

In these tests we used either the BTEC or Europarl corpus, and had only one phrase table, lexicalised reordering table, and language model. There is no adaptation in these cases, but each model is tuned to the SITAL data in devset5b-tune.

5.2. Corpus combination

The simplest way to combine two corpora is to append one to the other and train the system as if the data was from one corpus. Similar to single domain systems, there is still only one set of tables and language model. It is possible to do coarse weighting in this condition by duplicating the contents of one corpus multiple times within the file before training the system[9]. We conducted multiple experiments, incrementally increasing the number of copies of the smaller BTEC corpus from 1 to 8. We report our best results, which were obtained with a combination of six copies of BTEC and one copy of Europarl.

5.3. Separate corpora

In line with recent approaches to cross-domain adaptation, we take the components created from training each corpus separately and combine them at decoding time within one translation model. Using Moses' architecture, we can add a second language model and reordering model as additional, separate features in the model. In addition, we can use the multiple alternative decoding paths functionality to utilise multiple phrase tables[10]. We use two decoding paths (one for each corpus), each consisting of only one translation step.

Like our previous approach to domain adaptation[11], we maintain two separate phrase tables and language models. Unlike our previous work, we allow for two reordering tables, instead of one combined table as used in the previous system. While duplicating our previous best setup for separate language models and phrase tables, we tested using a BTEC lexicalised reordering table, a Europarl table, a combined table generated during the corpus combination experiment with one copy of each corpus, and finally, two separate lexicalised reordering tables.

5.4. Cross-domain adaptation results

The results of each of the approaches described above are presented in Table 4. We show BLEU scores for both 1-best and text input from devset5b-test. While having one combined reordering table and two separate reordering tables produced identical BLEU scores for the devset5b-test text data, the two table approach had a much better score on 1-best input.

6. Experiments in ASR input

The Moses MT system is capable of handling ASR input[12]. Previous work has shown that higher-scoring systems can be produced by using a confusion network than by using the 1-best ASR output as a text translation source. We investigated this in the case of the SITAL devset5b speech data, provided in SLF format.

For the results presented here in Table 5, we used devset4, devset5a and devset5b-tune as tuning sets, and devset4, devset5a and devset5b-test as development test sets. We tuned and tested on confusion network (CN), 1-best, and cleaned text input. For devsets 4 and 5a we did not split into -test and -tune subsets, so only results on the non-tuned set are shown. The first three experiments use the single BTEC cor-

Table 4: *Cross-domain adaptation results. For each model component, we use the BTEC corpus, the Europarl corpus, a concatenation of both in one training instance (Combined), or maintain two separate model components (BTEC, Europarl). Systems were tuned to devset5b-tune. Scores are shown in %*BLEU *for devset5b-test text and 1-best inputs.*

Method	Phrase table(s)	LM(s)	lexicalised reordering table(s)	%BLEU for TEXT	%BLEU for 1-BEST
5.1 Single corpus	Europarl	Europarl	Europarl	15.97	14.54
	BTEC	BTEC	BTEC	19.64	18.54
5.2 Corpus combination	Combined	Combined	Combined	21.51	20.43
5.3 Separate corpora	BTEC, Europarl	BTEC, Europarl	Europarl	21.54	19.68
	BTEC, Europarl	BTEC, Europarl	BTEC	22.92	20.82
	BTEC, Europarl	BTEC, Europarl	Combined	**23.02**	20.68
	BTEC, Europarl	BTEC, Europarl	BTEC, Europarl	**23.02**	**21.13**

pus translation model. The final setup uses the fully separate two corpus model described in section 5.3. Scores for devset4 and devset5a are higher at least in part due to multiple target reference translations for those sets, whereas devset5b-test has only single references.

As one would expect given previous work, the best ASR scores for devsets 4 and 5a are found on confusion network input using weights derived from confusion network tuning. However, we were unable to obtain satisfactory performance on the devset5b test set using confusion network input. The highest score for devset5b-test ASR input is always found in the 1-BEST column, not the CN column, and three of four times it is found from weights generated with text input tuning, not 1-best input.

Experiments in cross-domain adaptation for ASR were unsuccessful for confusion network tuning: due to time and memory restrictions we were unable to complete tuning runs for the full separate corpora setup that was our most successful system for cross-domain text translation adaptation.

7. Results

Based on the experiments above, we used the cross-domain adaptation weights with separate phrase and reordering tables and language models for each corpus for our official evaluation submission. Though we ran many experiments with the confusion network functionality of Moses, we were unable to satisfactorily tune the final system under confusion network input and had disappointing results with 1-best tuning. So, for the ASR track we submitted our text-tuned settings on 1-best input.

8. Conclusions

Our analysis of the SITAL data and experiments in cross-domain adaptation confirmed the benefits of using multiple corpora when translating from test sets without a direct in-domain training corpus, or with limited in-domain training data.

For the two development test sets within the BTEC domain, our ASR results were in agreement with previous work

Table 5: *ASR tuning results. All scores are shown in %*BLEU*. Best results for a given tuning/test set combination are shown in bold for ASR (CN or 1-BEST) and cleaned TEXT input conditions.*

BTEC CORPUS	**Test set devset4**		
Tuning set	**TEXT**	**1-BEST**	**CN**
devset5a TEXT	40.12	34.08	34.62
devset5a 1-BEST	40.43	34.31	34.41
devset5a CN	**40.99**	35.58	**36.08**
devset5b-tune TEXT	37.96	32.38	32.55
devset5b-tune 1-BEST	37.96	32.43	**32.65**
devset5b-tune CN	**38.15**	32.43	**32.65**
BTEC CORPUS	**Test set devset5a**		
Tuning set	**TEXT**	**1-BEST**	**CN**
devset4 TEXT	**37.53**	30.84	31.05
devset4 1-BEST	36.97	31.00	31.13
devset4 CN	37.15	31.02	**31.21**
devset5b-tune TEXT	34.53	29.14	29.10
devset5b-tune 1-BEST	34.93	29.47	29.61
devset5b-tune CN	**35.18**	29.53	**29.63**
BTEC CORPUS	**Test set devset5b-test**		
Tuning set	**TEXT**	**1-BEST**	**CN**
devset4 TEXT	**19.28**	**17.76**	17.37
devset4 1-BEST	17.69	16.04	15.83
devset4 CN	18.28	16.41	16.63
devset5a TEXT	15.70	14.62	13.78
devset5a 1-BEST	**19.55**	**17.91**	17.11
devset5a CN	18.03	16.51	16.13
devset5b-tune TEXT	19.64	**18.54**	18.42
devset5b-tune 1-BEST	19.52	18.38	18.12
devset5b-tune CN	**19.74**	18.24	17.94
SEPARATE CORPORA	**Test set devset5b-test**		
Tuning set	**TEXT**	**1-BEST**	**CN**
devset5b-tune TEXT	**23.02**	**21.13**	18.61
devset5b-tune 1-BEST	22.75	20.57	18.20
devset5b-tune CN	—	—	—

which suggests confusion network input improves translation performance over 1-best input. However, these results did not hold in our attempts to use the SITAL domain devset5b confusion network training data.

There are a number of possible causes for this result. At the data creation stage, it is possible that the ASR data is of a different quality for the SITAL input. During the tuning phase, the lack of multiple references for the SITAL data may have made tuning less effective. Final scoring may have been affected by the small number of sentences and single references, making it difficult to get accurate scores. It is also possible that the lack of an in-domain training corpus made it more difficult to translate the noisier ASR inputs.

In future work, we would like to improve the ability of the Moses system to dynamically cache phrase and reordering table entries for confusion network input. This would enable quicker confusion network decoding with a lower memory footprint than is currently possible, allowing us to scale to much larger training corpora. We would also like to test alternate methods for balancing and tuning multiple corpora within our system, scaling beyond two sets of training data.

9. Acknowledgements

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022 and in part under the EuroMatrix project funded by the European Commission (6th Framework Programme).

Thanks to Chris Dyer at the University of Maryland for scripts to process SLF data to Moses PCN format, and for assistance with our confusion network confusion.

10. References

[1] P. Koehn, "Europarl: A multilingual corpus for evaluation of machine translation," in *Proceedings of MT Summit*, 2005.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 177–180. [Online]. Available: http://www.aclweb.org/anthology/P/P07/P07-2045

[3] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase based 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)*, 2003.

[4] 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

[5] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh system description for the 2005 IWSLT speech translation evaluation," in *Proc. of the International Workshop on Spoken Language Translation*, October 2005.

[6] A. Arun, A. Axelrod, A. Birch Mayne, C. Callison-Burch, H. Hoang, P. Koehn, M. Osborne, and D. Talbot, "Edinburgh system description for the 2006 TC-STAR spoken language translation evaluation," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 37–41.

[7] A. Stolke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 2002.

[8] 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/W07-0717

[9] Y.-S. Lee, "IBM Arabic-to-English Translation for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 45–52.

[10] A. Birch, M. Osborne, and P. Koehn, "CCG supertags in factored statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 9–16. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0702

[11] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 224–227. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0733

[12] W. Shen, R. Zens, N. Bertoldi, and M. Federico, "The JHU Workshop 2006 IWSLT System," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 59–63.