Selecting Data for English-to-Czech Machine Translation *

Aleš Tamchyna, Petra Galuščáková, Amir Kamran, Miloš Stanojević, Ondřej Bojar
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Praha 1, CZ-118 00, Czech Republic
{tamchyna,galuscakova,kamran,bojar}@ufal.mff.cuni.cz,
milosh.stanojevic@gmail.com

Abstract

We provide a few insights on data selection for machine translation. We evaluate the quality of the new CzEng 1.0, a parallel data source used in WMT12. We describe a simple technique for reducing out-of-vocabulary rate after phrase extraction. We discuss the benefits of tuning towards multiple reference translations for English-Czech language pair. We introduce a novel approach to data selection by full-text indexing and search: we select sentences similar to the test set from a large monolingual corpus and explore several options of incorporating them in a machine translation system. We show that this method can improve translation quality. Finally, we describe our submitted system CU-TAMCH-BOJ.

1 Introduction

Selecting suitable data is important in all stages of creating an SMT system. For training, the data size plays an essential role, but the data should also be as clean as possible. The new CzEng 1.0 was prepared with the emphasis on data quality and we evaluate it against the previous version to show whether the effect for MT is positive.

Out-of-vocabulary rate is another problem related to data selection. We present a simple technique to reduce it by including words that became spurious OOVs during phrase extraction.

* This work was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Czech Republic) and the Czech Science Foundation grants P406/11/1499 and P406/10/P259.

Another topic we explore is to use multiple references for tuning to make the procedure more robust as suggested by Dyer et al. (2011). We evaluate this approach for translating from English into Czech.

The main focus of our paper however lies in presenting a method for data selection using full-text search. We index a large monolingual corpus and then extract sentences from it that are similar to the input sentences. We use these sentences in several ways: to create a new language model, a new phrase table and a tuning set. The method can be seen as a kind of domain adaptation. We show that it contributes positively to translation quality and we provide a thorough evaluation.

2 Data and Tools

2.1 Comparison of CzEng 1.0 and 0.9

As this year's WMT is the first to include the new version of CzEng (Bojar et al., 2012b), we carried out a few experiments to compare its suitability for MT with its predecessor, CzEng 0.9. Apart from size (which has almost doubled), there are important differences between the two versions. In CzEng 0.9, the largest portion by far came from movie subtitles (a data source of varying quality), followed by EU legislation and technical manuals. On the other hand, CzEng 1.0 has over 4 million sentence pairs from fiction and nearly the same amount of data from EU legislation. Roughly 3 million sentence pairs come from movie subtitles. This proportion of domains suggests a higher quality of data. Moreover, sentences in CzEng 1.0 were automatically filtered using a maximum entropy classifier that uti-

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 374–381,
Montréal, Canada, June 7-8, 2012.

Corpus and Domain		Sents	BLEU	Vocab. [k] En	Cs
CzEng 0.9	all	1M	14.77±0.12	187	360
CzEng 1.0			**15.23±0.18**	221	396
CzEng 0.9	news	100k	**14.34±0.05**	53	125
CzEng 1.0			14.01±0.13	47	113

Table 1: Comparison of CzEng 0.9 and 1.0.

lized a variety of features.

We trained contrastive phrase-based Moses SMT systems—the first one on 1 million randomly selected sentence pairs from CzEng 0.9, the other on the same amount of data from CzEng 1.0. Another contrastive pair of MT systems was based on small in-domain data only: 100k sentences from the *news* sections of CzEng 0.9 and 1.0. For each experiment, the random selection was done 5 times. In both experiments, identical data were used for the LM (News Crawl corpus from 2011), tuning (WMT10 test set) and evaluation (WMT11 test set).

Table 1 shows the results. The ± sign in this case denotes the standard deviation over the 5 experiments (each with a different random sample of training data). The results indicate that overall, CzEng 1.0 is a more suitable source of parallel data—most likely thanks to the more favorable distribution of domains. However in the small in-domain setting, using CzEng 0.9 data resulted in significantly higher BLEU scores.

The vocabulary size of the news section seems to have dropped since 0.9. We attribute this to the filtering: sentences with obscure words are hard to align so they are likely to be filtered out (the word alignment score as output by Giza++ received a large weight in the classifier training). These unusual words then do not appear in the vocabulary.

2.2 Lucene

Apache Lucene[1] is a high performance open-source search engine library written in Java. We use Lucene to take advantage of the information retrieval (IR) technique for domain adaptation. Each sentence of a large corpus is indexed as a separate document; a document is the unit of indexing and searching in Lucene. The sentences (documents) can then be retrieved based on Lucene similarity formula[2], given a "query corpus". Lucene uses Boolean model for initial filtering of documents. Vector Space Model with a refined version of Tf-idf statistic is then used to score the remaining candidates.

In the normal IR scenario, the query is usually small. However, for domain adaptation a query can be a whole corpus. Lucene does not allow such big queries. This problem is resolved by taking the query corpus sentence by sentence and searching many times. The final score of a sentence in the index is calculated as the average of the scores from the sentence-level queries. Methods that make use of this functionality are discussed in Section 5.

3 Reducing OOV by Relaxing Alignments

Out-of-vocabulary (OOV) rate has been shown to increase during phrase extraction (Bojar and Kos, 2010). This is due to unfortunate alignment of some words—no consistent phrase pair that includes them can be extracted. This issue can be partially overcome by adding translations of these "lost" words (according to Giza++ word alignment) to the extracted phrase table. This is not our original technique, it was suggested by Mermer and Saraclar (2011), though it is not included in the published abstract.

The extraction of phrases in the (hierarchical) decoder Jane (Stein et al., 2011) offers a range of similar heuristics. Tinsley et al. (2009) also observes gains when extending the set of phrases consistent with the word alignment by phrases consistent with aligned parses.

We evaluated this technique on two sets of training data—the news section of CzEng 1.0 and the whole CzEng 1.0. The OOV rate of the phrase table was reduced nearly to the corpus OOV rate in both cases, however the improvement was negligible—only a handful of the newly added words occurred in the test set. Table 2 shows the results. Translation performance using the improved phrase table was identical to the baseline.

[1] http://lucene.apache.org

[2] http://tiny.cc/ca2ccw

CzEng Sections	Test Set OOV % Baseline	Reduced	New Phrases
news (197k sents)	3.69	3.66	12034
all (14.8M sents)	1.09	1.09	154204

Table 2: Source-side phrase table OOV.

Sections	1 reference	3 references
news	11.37±0.47	**11.62±0.50**
all	**16.07±0.55**	15.90±0.57

Table 3: BLEU scores on WMT12 test set when tuning on WMT11 test set towards one or more references.

4 Tuning to Multiple Reference Translations

Tuning towards multiple reference translations has been shown to help translation quality, see Dyer et al. (2011) and the cited works. Thanks to the other references, more possible translations of each word are considered correct, as well as various orderings of words.

We tried two approaches: tuning to one true reference and one pseudo-reference, and tuning to multiple human-translated references.

For the first method, which resembles computer-generated references via paraphrasing as used in (Dyer et al., 2011), we created the pseudo-reference by translating the development set using TectoMT, a deep syntactic MT with rich linguistic processing implemented in the Treex platform[3]. We hoped that the very different output of this decoder would be beneficial for tuning, however we achieved no improvement at all.

For the second experiment we used 3 translations of WMT11 test set. One is the true reference distributed for the shared task and two were translated manually from the German version of the data into Czech. We achieved a small improvement in final BLEU score when training on a small data set. On the complete constrained training data for WMT12, there was no improvement—in fact, the BLEU score as evaluated on the WMT12 test set was negligibly lower. Table 3 summarizes our results. The ± sign denotes the confidence bounds estimated via bootstrap resampling (Koehn, 2004).

[3] http://ufal.ms.mff.cuni.cz/treex/

Used Models	Selected per Trans.	Sel. Sents Total	Avg BLEU±std
None	—	0	12.39±0.06
LM	—	16k – rand. sel.	12.18±0.06
LM	3	16k	12.73±0.04
LM	100	502k	14.21±0.11
LM	1000	3.8M	15.12±0.08
LM	All Sents	18.3M	**15.55±0.11**

Table 4: Results of experiments with Lucene, language model adapted.

5 Experiments with Domain Adaptation

Domain adaptation is widely recognized as a technique which can significantly improve translation quality (Wu et al., 2008; Bertoldi and Federico, 2009; Daumé and Jagarlamudi, 2011). In our experiments we tried to select sentences close to the source side of the test set and use them to improve the final translation.

The parallel data used in this section are only small: the news section of CzEng 1.0 (197k sentence pairs, 4.2M Czech words, 4.8M English words). We tuned the models on WMT09 test set and evaluated on WMT11 test set. The techniques examined here rely on a large monolingual corpus to select data from. We used all the monolingual data provided by the organizers of WMT11 (18.3M sentences, 316M words).

5.1 Tailoring the Language Model

Our first attempt was to tailor the language model to the test set. Our approach is similar to Zhao et al. (2004). In Moore and Lewis (2010), the authors compare several approaches to selecting data for LM and Axelrod et al. (2011) extend their ideas and apply them to MT.

Naturally, we only used the source side of the test set. First we translated the test set using a baseline translation system. Lucene indexer was then used to select sentences similar to the translated ones in the large target-side monolingual corpus. Finally, a new language model was created from the selected sentences.

The weight of the new LM has to reflect the importance of the language model during both MERT tuning as well as final application on (a different) test set. If the new LM were based only on the final

test set, MERT would underestimate its value and vice versa. Therefore, we actually translated both our development (WMT09) as well as final test set (WMT11) using the baseline model and created a LM relevant to their union.

The results of performed experiments with domain adaptation are in Table 4. To compensate for low stability of MERT, we ran the optimization five times and report the average BLEU achieved. The $\pm$ value indicates the standard deviation of the five runs.

The first row provides the scores for the baseline experiment with no tailored language model. We have run the experiment for three values of selected sentences per one sentence of the test corpus: 3, 100 and 1000 closest-matching sentences were extracted. With more and more data in the LM, the scores increase. The second line in Table 4 confirms the usefulness of the sentence selection. Picking the same amount of 16k sentences randomly performs worse. As the last row indicates, taking all available data leads to the best score.

Note that when selecting the sentences, we used lemmas instead of word forms to reduce data sparseness. So Lucene was actually indexing the lemmatized version of the monolingual data and the baseline translation translated English lemmas to Czech lemmas when creating the "query corpus". The final models were created from the original sentences, not their lemmatized versions.

5.2 Tailoring the Translation Model

Reverse self-training is a trick that allows to improve the translation model using (target-side) monolingual data and can lead to a performance improvement (Bojar and Tamchyna, 2011; Lambert et al., 2011).

In our scenario, we translated the selected sentences (in the opposite direction, i.e. from the target into the source language). Then we created a new translation model (in the original direction) based on the alignment of selected sentences and their reverse translation. This new model is finally combined with the baseline model and weighted by MERT. The whole scenario is shown in Figure 1.

The results of our experiments are in Table 5. We ran the experiment with translation model adaptation for 100 most similar sentences selected by Lucene. Each experiment was again performed five times. Due to the low stability of tuning, we also tried increasing the size of n-best lists used by MERT.

Experiments with tailored translation model are significantly better than the baseline but the improvement against the experiment with only the language model adapted (with the corresponding 100 sentences selected) is very small.

5.3 Discussion of Domain Adaptation Experiments

According to the results, using Lucene improves translation performance already in the case when only three sentences are selected for each translated sentence. Our results are further supported by the contrastive setup that used a language model created from a random selection of the same number of sentences—the translation quality even slightly degraded.

On the other hand, adding more sentences to language model further improves results and the best result is achieved when the language model is created using the whole monolingual corpus. This could have two reasons:

Too good domain match. The domain of the whole monolingual corpus is too close to the test corpus. Adding the whole monolingual corpus is thus the best option. For more diverse monolingual data, some domain-aware subsampling like our approach is likely to actually help.

Our style of retrieval. Our queries to Lucene represent sentences as simple bags of words. Lucene prefers less frequent words and the structure of the sentence is therefore often ignored. For example it prefers to retrieve sentences with the same proper name rather than sentences with similar phrases or longer expressions. This may not be the best option for language modelling.

Our method can thus be useful mainly in the case when the data available are too large to be processed as a whole. It can also highly reduce the computation power and time necessary to achieve good translation quality: the result achieved using the language model created via Lucene for 1000 selected sentences is not significantly worse than the result achieved using the whole monolingual corpus but the required data are 5 times smaller.

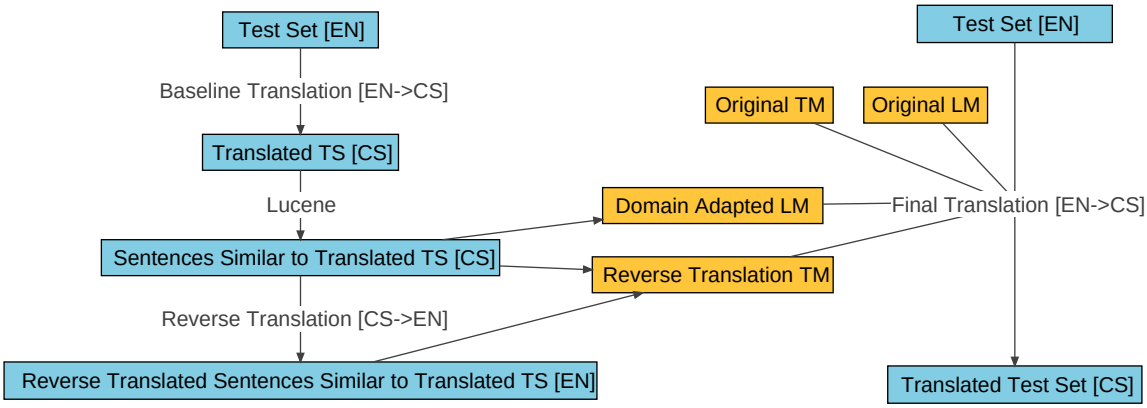

Figure 1: Scenario of reverse self-training.

Used Models	N-Best	Sel. Sents per Trans. Sent.	Sel. Sents Total	Avg BLEU±std
None	100	—	0	12.39±0.06
None	200	—	0	12.4±0.03
LM + TM	100	100	502k	14.32±0.13
LM + TM	200	100	502k	**14.36±0.07**

Table 5: Results of experiments with Lucene, translation model applied.

5.4 Tuning Towards Selected Data

Domain adaptation can also be done by selecting a suitable development corpus (Zheng et al., 2010; Li et al., 2004). The final model parameters depend on the domain of the development corpus. By choosing a development corpus that is close to our test set we might tune in the right direction. We implemented this adaptation by querying the source side of our large parallel corpus using the source side of the test corpus. After that, the development corpus is constructed from the selected sentences and their corresponding reference translations.

This experiment uses a fixed model based on the news section of CzEng 1.0. We only use different tuning sets and run the MERT optimization. All the resulting systems are tested on the WMT11 test set:

Baseline system is tuned on 2489 sentence pairs selected randomly from whole CzEng 1.0 parallel corpus. **Lucene** system uses 2489 sentence pairs selected from CzEng 1.0 using Lucene. The selection is done by choosing the most similar sentences to the source side of the final test set. **WMT10** system is tuned on 2489 sentence pairs of WMT10 test set. To identify an upper bound, we also include a **Perfect selection** system which is tuned on the final WMT11 test set. Naturally, this is not a fair competitor.

System	avg BLEU±std
Baseline	11.41±0.25
Lucene	12.31±0.01
WMT10	**12.37±0.02**
Perfect selection	12.64±0.02
Bad selection	6.37±0.64

Table 6: Results of tuning with different corpora

In order to make the results more reliable, it is necessary to repeat the experiment several times (Clark et al., 2011). Lucene and the WMT10 system were tuned 3 times while baseline system was tuned 9 times because of randomness in selection of tuning corpora (3 different tuning corpora each tuned 3 times). The results are shown in Table 6.

Even though the variance of the baseline system is high (because we randomly selected corpora 3

times), the difference in scores between baseline and Lucene system is high enough to conclude that tuning on Lucene-selected corpus helps translation quality. Still it does not give better BLEU score than system tuned on WMT10 corpus. One possible reason is that the whole CzEng 1.0 is of somewhat lower quality than the news section. Given that our final test set (WMT11) is also from the news domain, tuning towards WMT10 corpus probably leads to a better domain adaptation that tuning towards all the domains in CzEng.

The tuning set must not overlap with the training set. To illustrate the problem, we did a small experiment with the same settings as above and randomly selected 2489 sentences from training corpora. We again ran the random selection 3 times and tuned 3 times with each of the extracted tuning sets, see the "Bad selection" in Table 6.

In all the experiments with badly selected sentences, the distortion and language model get an extremely low weight compared to the weights of translation model. This is because they are not useful in translation of tuning data which was already seen during training. Instead of reordering two short phrases A and B, system already knows the translation of the phrase A B so no distortion is needed. On unseen sentences, such weights lead to poor results.

This amplifies a drawback of our approach: source texts have to be known prior to system tuning or even before phrase extraction.

There are methods available that could tackle this problem. Wuebker et al. (2010) store phrase pair counts per sentence when extracting phrases and thus they can reestimate the probabilities when a sentence has to be excluded from the phrase tables. For large parallel corpora, suffix arrays (Callison-Burch et al., 2005) have been used. Suffix arrays allow for a quick retrieval of relevant sentence pairs, the phrase extraction is postponed and performed on the fly for each input sentence. It is trivial to filter out sentences belonging to the tuning set during this delayed extraction. With dynamic suffix arrays (Levenberg et al., 2010), one could even simply remove the tuning sentences from the suffix array.

6 Submitted Systems

This paper covers the submissions CU-TAMCH-BOJ. We translated from English into Czech. Our setup was very similar to CU-BOJAR (Bojar et al., 2012a), but our primary submission is tuned on multiple reference translations as described in Section 4.

Apart from the additional references, this is a constrained setup. CzEng 1.0 were the only parallel data used in training. We used a factored model to translate the combination of English surface form and part-of-speech tag into Czech form+POS. We used separate 6-gram language models trained on CzEng 1.0 (interpolated by domain) and all News Crawl corpora (18.3M setences, interpolated by years). Additionaly, we created an 8-gram language model on target POS tags. For reordering, we employed a lexicalized model trained on CzEng 1.0.

Table 7 summarizes the official result of the primary submission and a contrastive baseline (tuned to just one reference translation). There is a slight decrease in BLEU, but the translation error rate (TER) is slightly better when more references were used. The differences are however very small, suggesting that tuning to more references did not have any significant effect.

System	BLEU	TER
multiple references	14.5	**0.765**
contrastive baseline	**14.6**	0.774

Table 7: Scores of the submitted systems.

7 Conclusion

We showed that CzEng 1.0 is of better overall quality than its predecessor. We described a technique for reducing phrase-table OOV rate, but achieved no improvement for WMT12. Similarly, tuning to multiple references did not prove very beneficial.

We introduced a couple of techniques that exploit full-text search in large corpora. We showed that adding selected sentences as an additional LM improves translations. Adding a new phrase table acquired via reverse self-training resulted only in small gains. Tuning to selected sentences resulted in a better system than tuning to a random set. However the Lucene-selected corpus fails to outperform good-quality in-domain tuning data.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 182–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ondřej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar and Aleš Tamchyna. 2011. Forms Wanted: Training SMT on Monolingual Data. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel, January.

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. 2012a. Probes in a Taxonomy of Factored Phrase-Based Models. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics. Submitted.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012b. The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC2012*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 255–262.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the Association for Computational Lingustics*. Association for Computational Linguistics.

Hal Daumé, III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 407–412, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Dyer, Kevin Gimpel, Jonathan H. Clark, and Noah A. Smith. 2011. The CMU-ARK German-English Translation System. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 337–343, Edinburgh, Scotland, July. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, Barcelona, Spain.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland, July. Association for Computational Linguistics.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 394–402.

Mu Li, Yinggong Zhao, Dongdong Zhang, and Ming Zhou. 2004. Adaptive development data selection for log-linear model in statistical machine translation. In *In Proceedings of COLING 2004*.

Coskun Mermer and Murat Saraclar. 2011. Unsupervised Turkish Morphological Segmentation for Statistical Machine Translation. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel, January.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Stein, David Vilar, Stephan Peitz, Markus Freitag, Matthias Huck, and Hermann Ney. 2011. A Guide to Jane, an Open Source Hierarchical Translation Toolkit. *Prague Bulletin of Mathematical Linguistics*, 95:5–18, March.

John Tinsley, Mary Hearne, and Andy Way. 2009. Exploiting parallel treebanks to improve phrase-based statistical machine translation. In Alexander F. Gelbukh, editor, *CICLing*, volume 5449 of *Lecture Notes in Computer Science*, pages 318–331. Springer.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 993–1000, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 475–484.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhongguang Zheng, Zhongjun He, Yao Meng, and Hao Yu. 2010. Domain adaptation for statistical machine translation in development corpus selection. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 2 –7, oct.