Data Selection With Fewer Words

Amittai Axelrod
University of Maryland
amittai@umd.edu

Philip Resnik
University of Maryland
resnik@umd.edu

Xiaodong He
Microsoft Research
xiaohe@microsoft.com

Mari Ostendorf
University of Washington
ostendo@uw.edu

Abstract

We present a method that improves data selection by combining a hybrid word/part-of-speech representation for corpora, with the idea of distinguishing between rare and frequent events. We validate our approach using data selection for machine translation, and show that it maintains or improves BLEU and TER translation scores while substantially improving vocabulary coverage and reducing data selection model size. Paradoxically, the coverage improvement is achieved by abstracting away over 97% of the total training corpus vocabulary using simple part-of-speech tags during the data selection process.

1 Introduction

Data selection uses a small set of domain-relevant data to select additional training items from a much larger, out-of-domain dataset. Its goal is to filter Big Data down to Good Data: finding the best, most relevant data to use to train a model for a particular task.

The prevalent data selection method, cross-entropy difference (Moore and Lewis, 2010), can produce domain-specific systems that are usually as good as or better than systems using all available training data (Axelrod et al., 2011). The size of these domain-specific systems scales roughly linearly with the amount of selected data: a system trained on the most domain-relevant 10% of the full out-of-domain dataset will be only one tenth of the size of a system trained using all the available data. This can be a large win in settings where training time matters, and also where compactness of the final system matters, e.g. running speech recognition or translation on mobile devices.

While data selection thus eliminates the need to train systems on the entire pool of available data, the data selection process itself does not scale well (it still requires a language model built on the entire pool) and, more significantly, it comes at a cost: training on selected subsets leads to reductions in vocabulary coverage compared to training on the full out-of-domain data pool. This coverage is important, because most NLP systems face the problem of handling words that were not seen in training the system, i.e. out-of-vocabulary (OOV) words. In automatic speech recognition (ASR), for example, OOV words pose a substantial problem, since the system will hallucinate a phonetically similar word in its vocabulary when an OOV word is encountered. In machine translation (MT), our focal application in this paper, OOVs can sometimes be transliterated, but often they are ignored or passed through without translation, and gaps in vocabulary coverage can have a significant effect on MT performance (Daumé III and Jagarlamudi, 2011; Irvine and Callison-burch, 2013).

We introduce a method that preserves the data selection benefit of reducing translation system size. Our method performs as well or better than the standard cross-entropy difference method, as measured by downstream MT results. To this we add the benefits of substantially improved lexical coverage, as well as lower memory requirements for the data selection model itself.

This improvement stems from constructing a hybrid representation of the text that abstracts away words that are infrequent in either of the in-domain and general corpora. They are replaced with their part-of-speech (POS) tags, permitting their n-gram statistics to be robustly aggregated: intuitively, if a domain-relevant sentence includes a rare word in some non-rare context (e.g. "An earthquake in Port-au-Prince"), then another sentence with the same context but a *different* rare word is probably also just as relevant (e.g. "An earthquake in Kodari"). While this method requires pre-processing the corpora to POS tag the

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 58–65,
Lisboa, Portugal, 17-18 September 2015.

data, the idea should generalize to automatically-derived word classes.

We present results using data selection to train domain-relevant SMT systems, yielding favorable performance compared against the standard approaches of Moore and Lewis (2010) and Axelrod et al. (2011). Paradoxically, this is achieved by a selection process in which the specific lexical items for infrequent words – up to 97% of the total vocabulary – are replaced with POS tags.

2 Related Work

Data selection is a widely-used variant of domain adaptation that requires quantifying the relevance to the domain of the sentences in a pooled corpus of additional data. The pool is sorted by relevance score, the highest ranked portion is kept, and the rest discarded.This process – also known as "rank-and-select" in language modeling (Sethy et al., 2009) – identifies the subset of the data pool that is most like the in-domain corpus and keeps it for translation system training, in lieu of using the entire data pool. The resulting translation systems are more compact and cheaper to train and run than the full-corpus system. The catch, of course, is that any large data pool can be expected to contain sentences that are at best irrelevant to the domain, and at worst detrimental: the goals of fidelity (matching in-domain data as closely as possible) and broad coverage are often at odds (Gascó et al., 2012). As a result, much work has focused on fidelity. Mirkin and Besacier (2014) survey the difficulties of increasing coverage while minimizing impact on model performance.

We build on the standard approach for data selection in language modeling, which uses *cross-entropy difference* as the similarity metric (Moore and Lewis, 2010). The Moore-Lewis procedure first trains an in-domain language model (LM) on the in-domain data, and another LM on the full pool of general data. It assigns to each full-pool sentence s a *cross-entropy difference score*,

$$H_{LM_{IN}}(s) - H_{LM_{POOL}}(s), \quad (1)$$

where $H_m(s)$ is the per-word cross entropy of s according to language model m. Lower scores for cross-entropy difference indicate more relevant sentences, i.e. those that are most like the target domain and unlike the full pool average. In bilingual settings, the *bilingual Moore-Lewis* criterion, introduced by Axelrod et al. (2011), combines the cross-entropy difference scores from each side of the corpus; i.e. for sentence pair $\langle s_1, s_2 \rangle$:

$$\begin{aligned}&(H_{LM_{IN_1}}(s_1) - H_{LM_{POOL_1}}(s_1)) \\ &+(H_{LM_{IN_2}}(s_2) - H_{LM_{POOL_2}}(s_2)) \quad (2)\end{aligned}$$

After sorting on the relevant criterion, the top-n sentences (or sentence pairs) are added to the in-domain data to create the new, combined training set. Typically a range of values for n is considered, selecting the n that performs best on held-out in-domain data.

While shown to be effective, however, word-based scores may not capture all facets of relevance. The strategy of a *hybrid* word/POS representation was first explored by Bulyko et al. (2003), who used class-dependent weights for mixing multi-source language models. The classes were a combination of the 100 most frequent words and POS tags. Bisazza and Federico (2012) target in-domain coverage by using a hybrid word/POS representation to train an additional LM for decoding in an MT pipeline. Toral (2013) uses a hybrid word/class representation for data selection for language modeling; he replaces all named entities with their type (e.g. 'person', 'organization'), and experiments with also lemmatizing the remaining words.

3 Our Approach: Abstracting Away Words in the Long Tail

Our approach is motivated by the observation that domain mismatches can have a strong register component, and this comprises both lexical and syntactic differences. We are inspired, as well, by work in stylometry, observing that attempts to quantify differences between text datasets try to learn too much from the long tail (Koppel et al., 2003): most words occur very rarely, meaning that empirical statistics for them are probably overestimating their seen contexts and underestimating unseen ones.

We therefore adopt a hybrid word/POS representation strategy, but, crucially, we focus not on restricting attention to *frequent* words, but on avoiding the undue effects of *infrequent* words. The proposal can be realized straightforwardly: after part-of-speech tagging the in-domain and pool corpora, we identify all words that appear infrequently in either one of the two corpora, and replace each of their word tokens with its POS tag.

Relevance computation, sentence ranking and subset selection then proceed as usual according to the Moore-Lewis or bilingual Moore-Lewis criterion.

As an example, consider again the phrases *"an earthquake in Port-au-Prince"* and *"an earthquake in Kodari"*, and suppose that the words *an*, *in*, and *earthquake* are above-threshold in frequency. Our hybrid word/POS representation for both sentences would be the same: *"an earthquake in NNP"*.

Our approach differs from the standard data selection method most significantly in its handling of rare words in frequent contexts. Consider a domain-specific n-gram context c that appears with a rare word w. For example, in a hypothetical news domain, let $c =$ *"an earthquake in"*, made up of common words, and let $w =$ *Port-au-Prince*. Suppose that the in-domain corpus contains the phrase *"an earthquake in Port-au-Prince"* eight times. The word w does not appear any other times in the in-domain corpus, and the word $w' =$ *Kodari* never appears at all.

Now suppose the out-of-domain pool corpus contains a sentence with *"an earthquake in Kodari"*. The standard Moore-Lewis method considers *Kodari* to be an unknown word, and so only credits that pool sentence with matching the elements of c. In contrast, our method replaces both rare words w and w' with their POS tag, *NNP*, so that the pool sentence contains *"an earthquake in NNP"*. Our method thus credits c from the in-domain corpus, like Moore-Lewis, but we *also* credit the sentence with matching the 4-gram *"an earthquake in NNP"*, which appears eight times in the in-domain corpus. Despite not appearing in the pool corpus, the rare word w from the in-domain corpus now provides us information about the relevance of pool sentences containing a syntactically similar rare word w' that shares the same context c.

4 Experimentation

We evaluate our data selection approach in a realistic small-in-domain-corpus setting, in two ways. First, as an intrinsic evaluation, we look at vocabulary coverage of the selected data relative to the in-domain training set, i.e. how many words from the in-domain corpus are out-of-vocabulary for selected data, since models trained on those data would not not be able to handle those words. Second, as an extrinsic evaluation, we use statistical machine translation as a downstream task.

4.1 Evaluation Framework

We define our in-domain corpus as the TED talk translations in the WIT^3 TED Chinese-English corpus (Cettolo et al., 2012), a good example of a subdomain with little available training data. We used the IWSLT *dev2010* and *test2010* sets (also from WIT^3) for tuning and evaluation. The larger pool from which we selected data was constructed from an aggregation of 47 LDC Chinese-English parallel datasets.[1] Table 1 contains the corpus statistics for the task and pool bilingual corpora.

Corpus	Sentences	Vocab (En)	Vocab (Zh)
TED (task)	145,901	49,323	64,616
LDC (pool)	6,025,295	458,570	714,628

Table 1: Chinese-English Parallel Data.

We used the KenLM toolkit (Heafield, 2011) to build all language models used in this work (i.e., both for data selection and for the MT systems used for extrinsic evaluation). In all cases the models were 4-gram LMs. We used the Stanford part-of-speech tagger (Toutanova et al., 2003) when constructing our hybrid representations, to generate the POS tags for each of the English and Chinese sides of the corpora.[2]

We consider three ways of applying data selection using the standard (fully lexicalized) corpus representation and our hybrid representation. The first two use the monolingual Moore-Lewis method (Equation 1) to respectively compute relevance scores using the English (output) side and the Chinese (input) side of the parallel corpora. The third uses bilingual Moore-Lewis (Equation 2) to compute the bilingual score over both sides.

Each of these three variants produces a version of the full pool in which the sentences are ranked by relevance score, from lowest score

[1] Specifically: LDC2000T47 LDC2002T01 LDC2003E07 LDC2003T17 LDC2004E12 LDC2004T07 LDC2005T06 LDC2006T04 LDC2007E101 LDC2007T09 LDC2007T23 LDC2008E40 LDC2008E56 LDC2008T06 LDC2008T08 LDC2008T18 LDC2009E16 LDC2009E95 LDC2009T02 LDC2009T06 LDC2009T15 LDC2010T03 LDC2010T10 LDC2010T11 LDC2010T12 LDC2010T14 LDC2010T17 LDC2010T21 LDC2012T16 LDC2012T20 LDC2012T24 LDC2013E119 LDC2013E125 LDC2013E132 LDC2013E83 LDC2013T03 LDC2013T05 LDC2013T07 LDC2013T11 LDC2013T16 LDC2014E08 LDC2014E111 LDC2014E50 LDC2014E69 LDC2014E99 LDC2014T04 LDC2014T11.

[2] The Stanford NLP tools use the Penn tagsets, which comprise 43 tags for English and 35 for Chinese.

	English	Chinese
TED vocab	49,323	64,616
LDC vocab	458,570	714,628
Joint vocab	470,154	729,283
LDC minus singletons	243,882	373,381
Baseline selection vocab	257,744	388,927

Table 2: Chinese and English vocabulary for the baseline selection process.

	English	Chinese
Joint vocab	470,154	729,283
Vocab with count ≥ 10	10,036	11,440
POS tags	42	35
Hybrid vocab	10,078	11,475

Table 3: Chinese and English vocabulary for the proposed selection process.

(most domain-like) to highest score (least domain-like). For each of those ranked pools, we consider increasingly larger subsets of the data: the best n = 50K, the best n = 100K, and so on. The largest subset we consider consists of the best n = 4M sentence pairs out of the 6M available.

4.1.1 Cross-Entropy Difference Baseline

In addition to comparing against a system trained on all the data, we compare against systems trained on data selected via the standard cross-entropy difference method. The joint vocabulary for the TED and LDC data is shown in Table 2. However, when training the language models used for the baseline selection process, we first pruned the singletons from the LDC vocabulary. This step is not necessary, but provides a slightly stronger baseline. The rationale is that ignoring LDC singletons avoids reserving too much probability mass for rare words outside of the domain of interest. Unlike the experimental systems below, pruning the lexicon simply ignores the words in the corpus and does not replace them with anything. This process removed 47% of the LDC vocabulary in each language. We then merged the remaining words from LDC with the complete TED lexicon. This produced a final vocabulary of 257,744 (En) and 388,927 (Zh) words for the baseline cross-entropy difference selection process, as shown in Table 2.

4.1.2 Hybrid Representation Systems

As our infrequent-word threshold (selected ahead of our experimentation), we retained words with a count of 10 or more in each corpus, and replaced all other words with their POS tags to create the hybrid corpus representation. The minimum count requirement reduced the vocabulary to 10,036 English words and 11,440 Chinese words, as shown in Table 3. All other words were replaced, thus a minimum count of 10 in each corpus eliminates over 97% of the vocabulary in each language. We previously found that setting the threshold to 10 is slightly better than a minimum count of 20 (Axelrod, 2014), and varying the threshold further is a topic for future work; see Section 5.

4.2 Results

4.2.1 Intrinsic Evaluation

As noted, each of the bilingual Moore-Lewis method and our hybrid word/POS variation produces a version of the additional training pool in which sentences are ranked by relevance. We then select increasingly larger slices of the data from 50k to 4M, as described in Section 4.1, and report results. As shown in Figures 1 and 2, the hybrid-selected models show consistently improved vocabulary coverage when compared head-to-head with models trained on data selected via a Moore-Lewis method, across all subsets. The only exception is when examining the vocabulary coverage in one language while selecting data based on the other one (*e.g.* selecting data using the English half but measuring the TED vocabulary coverage in Chinese), where our method provides only negligible improvement. Overall, the in-domain (TED) vocabulary coverage is up to 10% better with our proposed method, and the general-data (LDC) vocabulary coverage is up to 20% better.

Table 4 illustrates what this looks like in more detail for a single slice containing the top 2M sentence pairs. The table shows how many more vocabulary items are covered by the 2M sentence slice selected using our hybrid representation (the *Hyb* columns) than are covered by the best 2M sentences selected using the standard lexical representation (the *Std* columns).

Our method shows this improved vocabulary coverage regardless of whether one compares the vocabulary coverage of the methods on the English side (the first three rows) or the Chinese side (the second three rows) of the corpora. Furthermore, the results also hold regardless of which of the three ways of performing cross-entropy-

Lang	Method	TED Coverage Std	TED Coverage Hyb	LDC Coverage Std	LDC Coverage Hyb
En	Mono-en	67%	72%	42%	52%
	Mono-Zh	70%	71%	48%	54%
	Bilingual	68%	72%	42%	52%
Zh	Mono-En	70%	71%	38%	46%
	Mono-Zh	69%	73%	43%	62%
	Bilingual	69%	73%	37%	54%

Table 4: Vocabulary coverage comparison between standard and hybrid-based data selection, for data-selected samples of 2M sentences.

based data selection one uses. The three ways are: monolingual Moore-Lewis for the English and Chinese sides of the parallel corpus (*Mono-En* and *Mono-Zh*, respectively), as well as bilingual Moore-Lewis (*Bilingual*).

When selecting 2M sentences, Table 4 shows that the hybrid representation provides up to an extra 4-5% in-domain vocabulary coverage in either language. Furthermore, the hybrid-based methods obtain up to 10% more general-domain vocabulary coverage for English, and up to 19% more Chinese general-domain vocabulary coverage. All improvements are absolute percentage increases.

Figure 2 shows that our hybrid method's pool vocabulary coverage increases more rapidly than the baseline. The standard approach shows vocabulary coverage increasing more or less linearly with the amount of selection data. By contrast, our proposed method appears to asymptotically approach full in-domain vocabulary coverage, particularly for Chinese. Similarly, Figure 1 shows that our hybrid method also increases more rapidly to asymptotically approach full in-domain vocabulary coverage as well.

4.2.2 Extrinsic Evaluation

Improved vocabulary coverage is a positive result, but we are also interested in downstream application performance. Accordingly, we trained SMT systems using cdec (Dyer et al., 2010) on subsets of selected data. All SMT systems were tuned using MIRA (Chiang et al., 2008) on the dev2010 data from IWSLT (Federico et al., 2011), and then evaluated on the test2010 IWSLT test set using both BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). To isolate the impact of the data selection method, we present results just using the selected data, without the combining with the in-domain data into a multi-model system. Note that the hybrid word/POS representations were only used to compute the cross-entropy difference scores for determining sentences' relevance; the MT systems themselves are trained using the sentences containing the original words.

Figure 3 shows our MT results using both BLEU and TER. The horizontal line is a static baseline that uses all the available training data without data selection. The dashed grey line is from systems trained on data selected via the standard Moore-Lewis cross-entropy-difference method, and the black line is from systems trained on data selected with our hybrid approach. To account for variability in MT tuning, each of the curves in Figure 3 is the average of three tuning/decoding runs.

In terms of system accuracy, our results confirm prior work on data selection, demonstrating that in comparison to training using all available data, comparable or even better MT performance can be obtained using only a fraction of the out-of-domain data available.

Table 5 shows SMT results for the same subset size of 2M sentences used for the coverage results in Table 4. Systems trained on data selected using the hybrid representation are up to +0.5 BLEU better, regardless of whether the selection process is monolingual or bilingual. Indeed, at least for BLEU, it appears that our hybrid method may tend to converge to comparable performance more quickly, a possibility worthy of future experimentation.

The TER results are mixed for this data selection subset size. The MT evaluation scores are low in absolute terms, due to only using the general-domain data, yet are still not inconsistent with prior research done using this dataset (Federico et al., 2011). Fluctuations in the performance curves are also consistent with prior work, as IWSLT scores are very jittery. We averaged results over three tuning runs, for stability. Despite that, Figure 3 shows how high-variance TER scores are on this task.

4.2.3 Selection Model Size

The resulting translation system sizes conform with prior work: selecting smaller subsets yields smaller downstream MT systems. For example, an MT system trained on 1M selected sentences is $\sim$2.3GB in size, a factor of 5 smaller than the 11.3GB baseline MT system trained on all 6M sentences. In addition, we observe a healthy re-

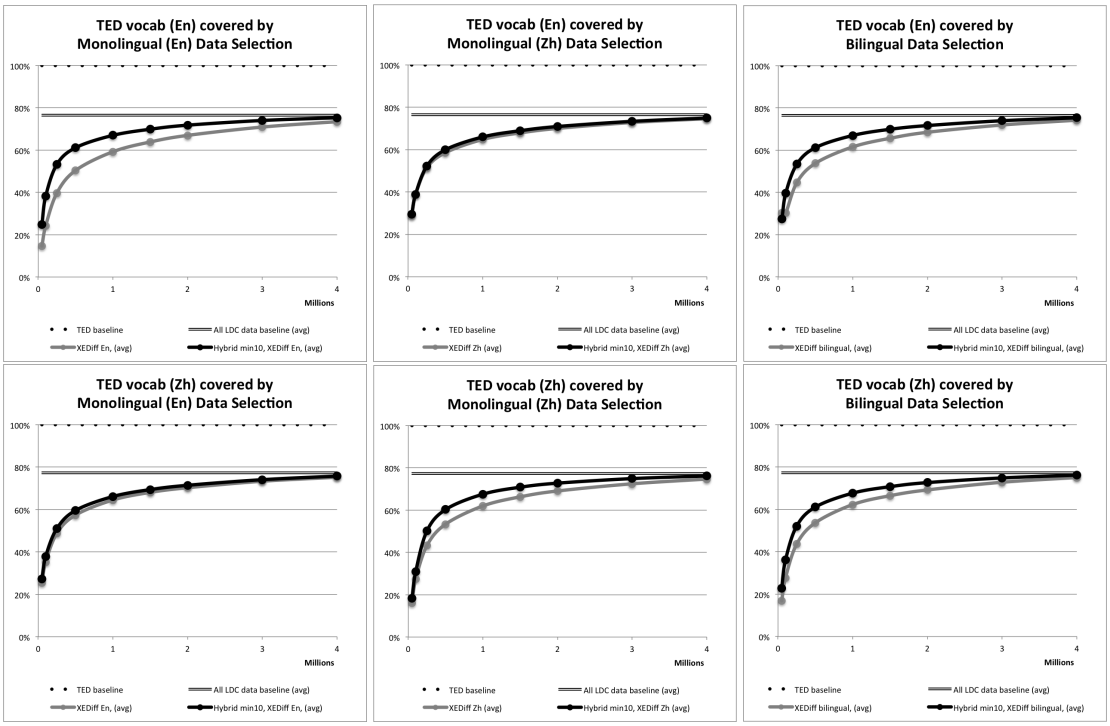

Figure 1: Percentage of TED vocabulary covered vs. number of selected sentences by method.

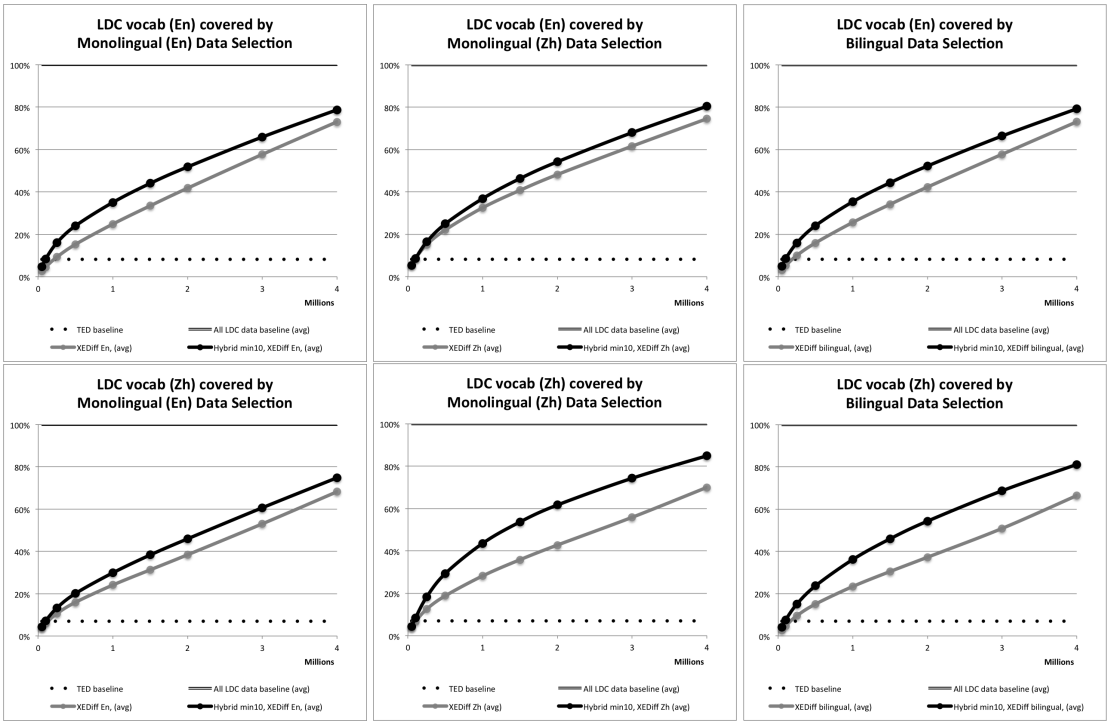

Figure 2: Percentage of LDC vocabulary covered vs. number of selected sentences by method.

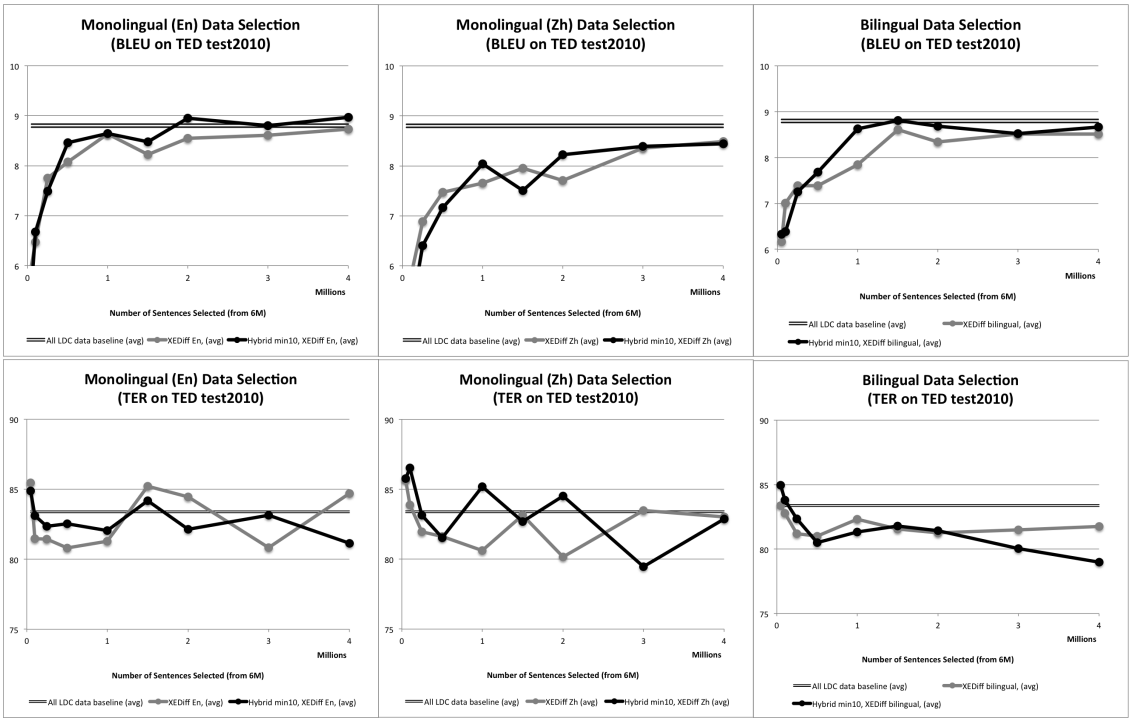

Figure 3: SMT system scores on the TED Zh-En test2010 set vs. number of selected sentences by method.

Metric	Method	Std	Hyb
BLEU	Mono-en	8.55	8.95
	Mono-Zh	7.70	8.22
	Bilingual	8.34	8.68
TER	Mono-En	84.44	82.15
	Mono-Zh	80.16	84.51
	Bilingual	81.27	81.44

Table 5: SMT system score comparison between standard and hybrid-based data selection, for data-selected samples of 2M sentences.

duction in the memory requirements for the data selection process, which requires training a language model on the entire data pool. The binarized language model built using the standard data selection baseline on the full corpus of 6M sentences requires about 2GB, whereas the equivalent all-data LM for our approach is 25% smaller.[3] This means that for any given amount of available memory, the hybrid method can scale up data selection to a larger out-of-domain sentence pool. As a rough example, an 8GB desktop machine can be used to train an LM on 32M sentences using the hybrid representation rather than 24M using the standard text; for a large-memory 128GB machine, our method would allow us to increase the size of the corpus used to train the full-data LM from a maximum of 384M sentences to more than half a billion sentences.

[3]Our back-of-the-envelope estimates ignore the in-domain LM, which is tiny in comparison.

5 Conclusions

We have presented a new method for data selection that retains the existing advantages of the state-of-the-art approach, while improving vocabulary coverage and also improving the ability to scale up to larger out-of-domain datasets. Our motivation is in the practical application of NLP technology, which often requires working with constrained resources and in specific domains with limited training data. The proposal is conceptually simple, uses widely available tools, and is easily applied. A drawback of the proposed approach is that it requires an additional pre-processing step of tagging all of the training data. For languages for which a POS tagger is not available, we expect that data-driven word classes would be a good substitute. In future work we plan to explore hybrid representations further, e.g. abstracting away from infrequent lexical items via distributional clustering or morphological analysis, rather than using part-of-speech information.

Acknowledgments

We gratefully thank the anonymous reviewers and Timo Baumann for their detailed feedback.

References

Axelrod, A. (2014). *Data Selection for Statistical Machine Translation*. PhD thesis, University of Washington.

Axelrod, A., He, X., and Gao, J. (2011). Domain Adaptation Via Pseudo In-Domain Data Selection. *EMNLP (Empirical Methods in Natural Language Processing)*.

Bisazza, A. and Federico, M. (2012). Cutting the Long Tail : Hybrid Language Models for Translation Style Adaptation. *EACL (European Association for Computational Linguistics)*, pages 439–448.

Bulyko, I., Ostendorf, M., and Stolcke, A. (2003). Getting More Mileage From Web Text Sources For Conversational Speech Language Modeling Using Class-Dependent Mixtures. *NAACL (North American Association for Computational Linguistics)*.

Cettolo, M., Girardi, C., and Federico, M. (2012). WITˆ3 : Web Inventory of Transcribed and Translated Talks. *EAMT (European Association for Machine Translation)*.

Chiang, D., Marton, Y., and Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. *EMNLP (Empirical Methods in Natural Language Processing)*.

Daumé III, H. and Jagarlamudi, J. (2011). Domain Adaptation for Machine Translation by Mining Unseen Words University of Maryland. *ACL (Association for Computational Linguistics)*.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blumson, P., Setiawan, H., Eidelman, V., and Resnik, P. (2010). cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models. *ACL (Association for Computational Linguistics) Interactive Poster and Demonstration Sessions*, (July):7–12.

Federico, M., Bentivogli, L., Paul, M., and Stüker, S. (2011). Overview of the IWSLT 2011 Evaluation Campaign. *IWSLT (International Workshop on Spoken Language Translation)*.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does More Data Always Yield Better Translations? *EACL (European Association for Computational Linguistics)*.

Heafield, K. (2011). KenLM : Faster and Smaller Language Model Queries. *WMT (Workshop on Statistical Machine Translation)*.

Irvine, A. and Callison-burch, C. (2013). Combining Bilingual and Comparable Corpora for Low Resource Machine Translation. *WMT (Workshop on Statistical Machine Translation)*.

Koppel, M., Akiva, N., and Dagan, I. (2003). A Corpus-Independent Feature Set for Style-Based Text Categorization. *IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis*.

Mirkin, S. and Besacier, L. (2014). Data Selection for Compact Adapted SMT Models. *AMTA (Association for Machine Translation in the Americas)*.

Moore, R. C. and Lewis, W. D. (2010). Intelligent Selection of Language Model Training Data. *ACL (Association for Computational Linguistics)*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-j. (2002). BLEU: a method for automatic evaluation of machine translation. *ACL (Association for Computational Linguistics)*.

Sethy, A., Georgiou, P. G., Ramabhadran, B., and Narayanan, S. S. (2009). An iterative relative entropy minimization based data selection approach for n-gram model adaptation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):13–23.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *AMTA (Association for Machine Translation in the Americas)*, (August):223–231.

Toral, A. (2013). Hybrid Selection of Language Model Training Data Using Linguistic Information and Perplexity. *Workshop on Hybrid Approaches to Translation*, pages 8–12.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *NAACL (North American Association for Computational Linguistics)*.