Class-Based N-gram Language Difference Models for Data Selection

Amittai Axelrod
Johns Hopkins University
and University of Maryland
amittai@clsp.jhu.edu

Yogarshi Vyas, Marianna Martindale, Marine Carpuat
University of Maryland
marine@cs.umd.edu

Abstract

We present a simple method for representing text that explicitly encodes differences between two corpora in a domain adaptation or data selection scenario. We do this by replacing every word in the corpora with its part-of-speech tag plus a suffix that indicates the relative bias of the word, or how much likelier it is to be in the task corpus versus the pool. By changing the representation of the text, we can use basic n-gram models to create *language difference models* that characterize the difference between the corpora. This process enables us to use common models with robust statistics that are tailored to computing the similarity score via cross-entropy difference.

These improvements come despite using zero of the original words in the texts during our selection process. We replace the entire vocabulary during the selection process from 3.6M to under 200 automatically-derived tags, greatly reducing the model size for selection.

When used to select data for machine translation systems, our language difference models lead to MT system improvements of up to +1.8 BLEU when used in isolation, and up to +1.3 BLEU when used in a multi-model translation system. Language models trained on data selected with our method have 35% fewer OOV's on the task data than the most common approach. These LMs also have a lower perplexity on in-domain data than the baselines.

1. Introduction

Data selection is a popular approach to 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 of the data discarded. By identifying the subset of the data pool that is most like the in-domain corpus and using it instead of the entire data pool, the resulting translation systems are more compact and cheaper to train and run than the full system trained on all of the available data. The underlying assumption in data selection is that the large corpus likely includes some sentences that fall within the target domain. These in-domain sentences should be used for training. Any large data pool will also contain sentences that are irrelevant at best to the domain of interest. At worse, these sentences that are *so* unlike the in-domain data that their presence makes the downstream models worse, and thus they should be removed from the training set.

We note that the models used for data selection are n-gram language models. These are typically used to characterize an entire corpus. However, the data selection scenario is not a characterization task, but a differentiating one. For every sentence in some large, general data pool of potentially dubious provenance, we would like to compute its relevance to some particular in-domain corpus, regardless of what it contains. One could even claim that we do not care what the in-domain data looks like, we just want more of whatever it is.

This supports the use of different models for selecting the data than for using the data in some downstream application. In particular, during the selection process it is more important to know how the corpora differ than how they are alike. We present a simple method for constructing a discriminative representation of the general corpus, and use it to train a language model that is focused on quantifying the difference between the in-domain and general corpora.

2. Background

2.1. Data Selection

The standard approach for data selection uses *cross-entropy difference* as the similarity metric [1]. This procedure leverages the mismatch between the data pool

Proceedings of the 12th International Workshop on Spoken Language Translation
Da Nang, Vietnam, December 3-4, 2015

and the task domain. It first trains an in-domain language model (LM) on the task 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 task *and most unlike* the full pool average. In bilingual settings such as machine translation, the *bilingual Moore-Lewis* criterion [2] 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 selected to create a task-relevant training set. Typically a range of values for n is considered, selecting the n that performs best on held-out in-domain data.

Cross-entropy difference data selection methods are a common pre-processing step for machine translation applications where model size or domain specificity are important. These methods have been extended within the MT community, *e.g.* by [3] using IBM model socres, edit distance [4], neural language models [5]. Furthermore, [6] showed improvements by using EM to identify true out-of-domain data from the pool to contrast against the in-domain data. They also highlight the distrinction between *relevance* and *fluency* that underlies the proposed language difference models. More recently, [7] proposed abstracting away rare words while training the models used for the selection step.

We present a simple method for modeling the difference between two corpora, one that is tailored to fit existing cross-entropy methods for data selection and can readily be applied to other problems.

2.2. Some Words Matter More

All words in a text do not contribute equally to characterize the text. However, which words are more important than others depends on the application. The most frequent words get higher probability in a normal n-gram language model. In topic modeling, content words are prized for what they convey and stopwords are ignored. By contrast, content words are largely ignored in stylometry when deciding the relevance of a text collection to a particular author or genre. Instead, the relevance is determined using function word and part of speech features together. In particular, [8] uses the difference in word frequencies across authors, genres, or eras. Syntactic structure or at least certain syntactic constructions are a potentially more informative source of stylometric features, [9] and [10]. POS tag sequences were introduced as stylometric features by [11] for document classification. [12] subsequently noted that the frequency of the word should be taken into account, else the classifier learns too much about rare events whose empirical estimates of counts and contexts might be incomplete.

A common thread is abstracting words into classes or groups that have more robust statistics. Sequences of these classes, such as part-of-speech (POS) tags, are then used as lightweight representations of the syntactic structure of a sentence. These can be thought of as a quantifiable proxy for sentence register, style, genre, and other ways of characterizing a corpus. For the more specific task of domain adaptation or data selection, replacing some words in the text with their POS tags is a way of creating general templates of what the text is like. This has been used in MT to build better domain-adapted language models [13] and for broader-coverage data selection [7] as mentioned previously.

3. Proposed Method

The method of [1] for data selection explicitly takes advantage of the inherent difference between the task and the pool corpora. Looking for sentences that are like the task corpus and are unlike the pool does not work if the two corpora are very similar. The language models trained on similar corpora will have similar distributions, so the scores in Equation 1 will subtract to zero.

However, in a domain adaptation scenario, the existence of a substantial difference between the task and pool corpora is axiomatic. If this were not the case, then there would be no adaptation scenario! The cross-entropy difference method exploits this difference between the corpora. Because the corpora must differ, then so must be the language models trained on them. Because the language models must differ, then subtracting the scores finds more relevant sentences.

We perform a similar trick with the text itself: where there is a difference between the language mod-

els trained on the task and the pool, then there is a difference between the frequencies of certain words in the corpora. Where the frequencies of words differ, the corpora differ. Where they do not differ, neither do the corpora, so we can expect to see them at the same rate. We can exploit this difference, because we know we are going to subtract the cross-entropy scores.

Words that appear with approximately the same frequency in both texts will have roughly similar cross-entropies according to both the task and pool language model. These words contribute negligibly to the cross-entropy difference scoring because the Moore-Lewis criterion subtracts the two language model scores. This means that the similarity score is only based on words whose empirical distributions are substantially different from one corpus to the other. These words appear in n-grams whose probabilities also differ between the corpora, and these are the non-zero components of the cross-entropy difference score for the sentence.

Whether the word is common or rare or inherently topical has little bearing on the score: if it appears similarly often in both corpora – regardless of how often that is – it will not contribute to the cross-entropy difference. A word's impact on data selection depends on the two corpora being compared in a specific data selection or domain adaptation scenario.

We can take advantage of this to construct models of the corpora that specifically capture which words matter for computing cross-entropy difference between these specific two in-domain and pool data sets. Rather than build new infrastructure, we will simply construct a representation of the text that captures this discriminative information, and then train an n-gram language model on the new representation. This approach has the advantage of being readily reproducible. We call the resulting model a *l*anguage difference model, and use it to compute the cross-entropy difference scores.

The representation of the text is straightforward: we replace each and every word with a token consisting of two parts: the POS tag of the word, and a suffix indicating how much more likely the word is to appear in the task corpus than in the pool corpus.

We use the *ratio of the word's probabilities* in the corpora to determine how much the two specific corpora differ with respect to a word. The ratio simply divides the frequency of the word in the task corpus by the frequency of the word in the pool corpus. This can also be readily computed using unigram LMs trained on each of the corpora.

In this particular work we distinguish this ratio as being quantized by powers of ten, as shown in Table 1. We also add an eighth suffix ("/low") to indicate words that occur fewer than 10 times, following the results in [7]. This was done to enable direct comparison of the contribution of the skew suffixes with prior work. In general, we only bucketed the probability ratios by powers of ten to demonstrate the potential of language difference models for data selection. There is ample room for exploration.

Frequency Ratio ($\frac{Task}{Pool}$)	Suffix	Example Token
$1000 \leq x$	/+++	JJ/+++
$100 \leq x < 1000$	/++	NNS/++
$10 \leq x < 100$	/+	NN/+
$10^{-1} \leq x < 10$	/0	DET/0
$10^{-2} \leq x < 10^{-1}$	/-	NN/-
$10^{-3} \leq x < 10^{-2}$	/- -	JJ/- -
$x < 10^{-3}$	/- - -	NNP/- - -

Table 1: Suffixes to indicate how indicative a word is of one corpus or the other

Our class-based n-gram language difference model representation condenses the entire vocabulary from hundreds of thousands of words down to 150-190 total types, as shown in Table 2. Each type conveys a class of words' syntactic information –which can be considered a proxy for style – as well as information about how indicative the words are of one corpus or the other.

Language	Vocab (full)	Labels (Task)	Labels (Pool)
English	3,904,187	148	182
French	3,681,086	147	190

Table 2: Corpus vocabulary size before and after replacing all words with discriminative labels

As an example, consider the word *supermassive*, which appears 21 times in the in-domain corpus, and 35 times in the data pool. The task pool contains 4.2M tokens, and the data pool contains 1,180M tokens. The empirical frequency ratio:

$$\frac{C_{task}(supermassive)}{4.2M} \div \frac{C_{pool}(supermassive)}{1,180M}$$

is calculated by:

$$\frac{1,180M}{4.2M} \times \frac{C_{task}(supermassive)}{C_{pool}(supermassive)} \approx 281 * \frac{21}{35} = 169$$

The derivation of the labels used to replace a phrase such as *supermassive black holes* in the class-based language difference representation used for data selection is shown in Table 3.

words:	supermassive	black	holes
POS:	JJ	JJ	NNS
ratio:	$100 \leq 169 < 1000$	$10^{-1} \leq 8 < 10$	$10 \leq 28 < 100$
label:	JJ/++	JJ/0	NNS/+

Table 3: Deriving the discriminative representation of a phrase. Only the tokens in the last line appear in the language difference model, as they are 1-to-1 replacements for the original words in first line.

Once the text has been transformed into the class-based language difference representation, we proceed with the standard cross-entropy difference algorithm. After computing the similarity scores and using them to re-rank the sentences in the pool corpus, we transform the text back into the original words and train the downstream LMs and SMT systems as normal. This process enables us to use models with robust statistics for how the corpora differ in order to compute the relevance score, and then use the traditional, n-gram based systems for the downstream MT pipeline.

4. Experimental Framework

Our experiments were based on the French-to-English MT evaluation track for IWSLT 2015. The task domain was defined to be TED talks, a translation subdomain with only 207k parallel training sentences. The data pool consisted of 41.3M parallel sentences from assorted sources, described in Table 4. The parallel Wikipedia and TED corpus were from the ISWLT 2015 website.[1] The remaining corpora were obtained from WMT 2015.[2] Our systems were tuned on *test2010* and evaluated using BLEU [14] on *test2012*, and *test2013* from the same TED source.

All parallel data was tokenized with the Europarl tokenizer[3] and lowercased with the `cdec` tool. We found it was necessary to further preprocess the data by using perl's Encode module to encode as UTF-8 octets and decode back to characters. We replaced fatally malformed characters with the Unicode replacement character, `U+FFFD`.

Dataset	# of sentences
Europarl v7	2.0M
News Commentary	0.2M
Common Crawl	3.2M
10^9 Fr-En	22.5M
UN Corpus	12.8M
Wikipedia	0.4M
TED corpus	0.2M

Table 4: Provenance of the 41M sentence French-English data pool.

We trained all SMT systems using cdec [15], tuned with MIRA [16]. The (4-gram) language models used for the selection process were all trained with KenLM [17]. The Stanford part-of-speech tagger [18] generated the POS tags for both English and French.[4]

5. Results and Discussion

The standard Moore-Lewis data selection method uses normal n-gram language models to compute the cross-entropy scores according to each of the task and pool language models. These scores get subtracted into the cross-entropy difference score that is used to rank each sentence in the data pool. We have proposed computing these cross-entropy values differently: using a language model trained over the class-based language difference labels for each word in the sentence, instead of the LM trained on the words themselves.

As an experimental baseline, we perform Moore-Lewis data selection in the standard way using the normal text corpora (`"xediff"`). This method is shown in grey in all figures. The language models were standard n-gram word-based models, with order 4 and the vocabulary fixed to be the pool lexicon minus singletons, plus the task lexicon. The final English-side vocabulary contained 1,796,862 words, and the French side 1,728,231, with the size reflecting the noisiness and

[1] `https://sites.google.com/site/iwsltevaluation2015/data-provided`

[2] `http://www.statmt.org/wmt15/translation-task.html`

[3] `http://www.statmt.org/europarl/v7/tools.tgz`

[4] The Stanford NLP tools use the Penn tagsets, which comprise 43 tags for English and 31 for French.

heterogeneity of the data pool.

For a slightly harder baseline, we compare against the approach of [7] in WMT 2015, who replace all rare words (count < 10) with their POS tag during the selection process (`"min10"`). This method is shown in dark blue in all figures. This baseline is expected to provide a modest improvement in translation quality and a large improvement in lexical coverage. Finally we perform our proposed method of language difference models, replacing all words in the corpora with a class-based difference representation during the selection process (`"new"`). These results are shown in orange.

Each of these variants produces a version of the full pool in which the sentences are ranked by relevance score. For each of those ranked pools, we evaluate language models trained on increasingly larger *slices* of the data ranging from the highest scoring $n = 500$K to the highest scoring $n = 5$M sentence pairs out of the 41M available. We performed all experiments three times: using only monolingual score on each language, and using the bilingual score. We report only results on monolingual English method due to space; the trends were the same in all tracks.

5.1. Language Modeling

Figure 1 shows language modeling results. We present results only for monolingual English-side data selection, but the results for monolingual French-side and bilingual data selection are similar. For each of the three data selection methods, we trained language models on the most relevant subsets of various sizes. The language models were configured identically to those used for selection (order 4, and vocabulary fixed).

We evaluated these models on their perplexity on the entire TED training set (207k sentences). The recent work of [7] `"min10"` does not beat the vanilla Moore-Lewis baseline perplexity, although they converge. On the left side of Figure 1, it can be seen that proposed method of language difference models provides a clear and consistent reduction of 13 perplexity (absolute; 10% relative) over the standard word-based method. This is roughly the same perplexity improvement as was shown in [1], so adding the discriminative information to the text doubles the effectiveness of cross-entropy difference -based data selection.

The right-hand side of Figure 1 shows the number of out-of-vocabulary (OOV) tokens in the TED task corpus according to LMs trained on the selected data. We confirm the large vocabulary coverage improvement reported in [7], with `"min10"` having 43% (relative) fewer OOV's at the 2-3M selection mark. Our proposed new method is almost as good, with 37% fewer OOV's on the task.

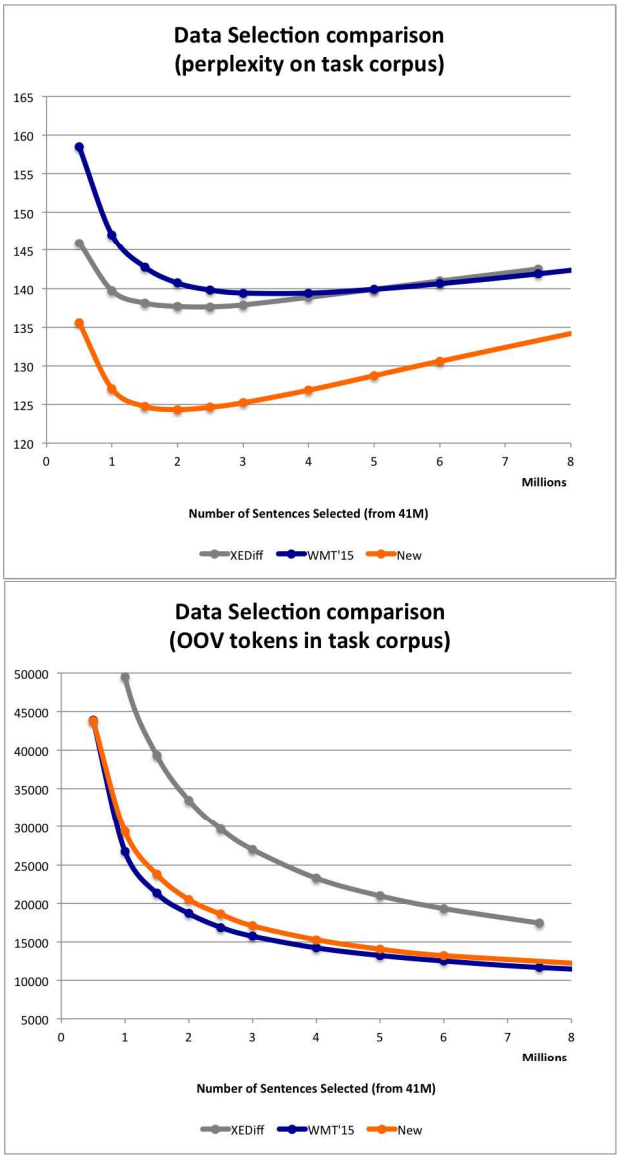

Figure 1: Comparison of perplexity scores and OOV tokens on the TED corpus for monolingual (English) data selection with word only, words-and-POS, and with language difference information.

5.2. Machine Translation

The machine translation results comparing textual representations for each data selection variant are in Figure 2. The BLEU scores of systems from the class-based language difference (`"new"`) approach are all

substantially better than either baseline on both tst2013 and tst2012.

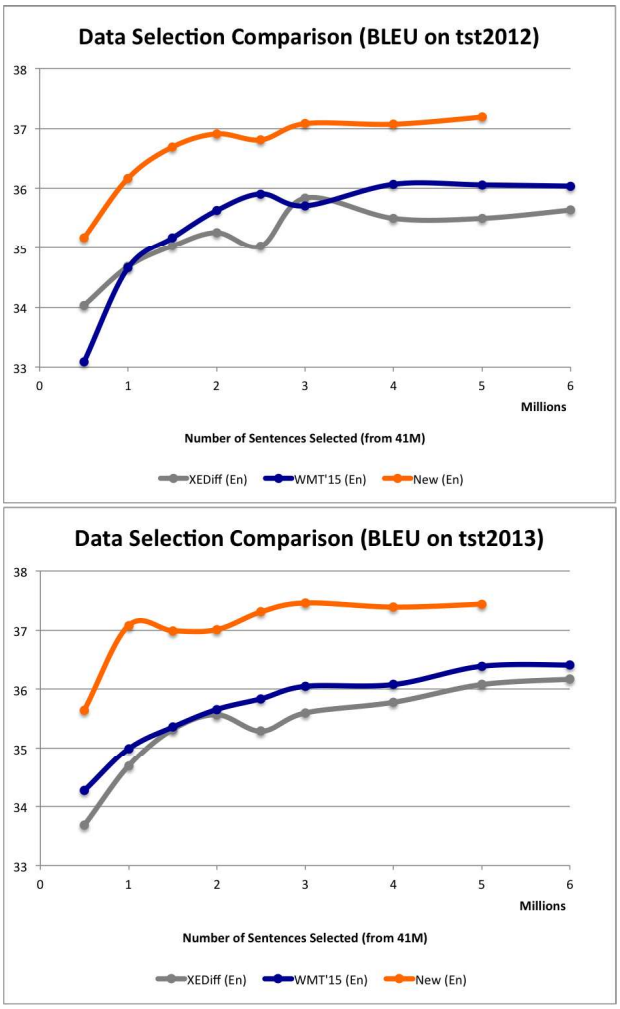

Figure 2: Comparison of BLEU scores for monolingual (English) data selection with word only, words-and-POS, and with language difference information.

When selecting 3M selected sentences and evaluating on the most recent public test set, `test2013`, the Moore-Lewis baseline of [1] has a BLEU score of 35.60, the fewer-words (”min10”) baseline from [7] scores 36.05 (+0.45), and our new method scores 37.46, +1.85 BLEU over the first baseline and +1.4 over the second, more recent, update to the state-of-the-art. The BLEU scores of the proposed method reach a higher plateau, and do so earlier. Of note is that only the language difference models select data that outperforms the in-domain corpus (the black line labeled “TED baseline” in Figure 3.

We also tested using the selected data to build a multi-model system, where the translation model trained on selected data is used in combination with one trained on the task data. Each resulting system thus had two grammars and two language models. Figure 3 contains the results of these multi-model experiments using the monolingual (English) selection method, and evaluated on “test2013”. All the data selection methods provided some benefit when used in the multi-model setup, but the proposed method using language difference models was up to +1 BLEU better than the baseline in [7] (which did not show multi-model results), up to +1.3 BLEU than the cross-entropy difference baseline, and +2 BLEU over the in-domain data alone.

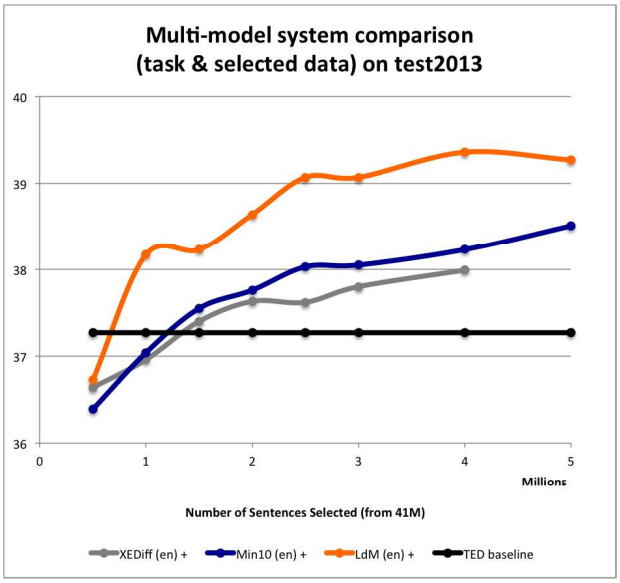

Figure 3: Using system trained on selected data as part of a two-model translation system, along with a system trained on the task corpus.

It is thus possible to use a wordless text representation to select data more usefully than a word-based method. This is surprising to us, as the language models trained on our class-based language difference text have no way of knowing if the sentences being scored are topically relevant. Modeling the difference between corpora in aggregate can thus be a stronger indicator of relevance than the words themselves for selection. We collapsed all of the words in the vocabulary as a pathological test case; a more finely-tuned approach would perhaps distinguish between words to keep and words to abstract away into a difference class.

5.3. Model Size Improvements

In addition to the translation system improvements, the memory requirements for the data selection pro-

cess itself dominated by the language model built using the data pool is dramatically smaller with our class-based n-gram language difference representation than the baseline models. The standard data selection method requires training a 12GB (binarized) language model over each side of the full 42M sentence pool in order to compute the cross-entropy score according to the general-domain corpus. The equivalent full-corpus model using our approach is 126M, or 1% as large, because the vocabulary size is negligible.

5.4. Requirements

One drawback to this use of language difference models as presented here is our use of a part-of-speech tagger in at least one of the languages. Languages with large amounts of data generally seem to have POS taggers already developed. However, there are plenty of languages for which such linguistic tools are not accessible. To construct the language difference model, the discriminative (or skew) information about each word is combined with some generalization or group label for the word that conveys part of the word's information in the sentence. POS tags are just one of many ways of grouping words together so as to capture underlying relationships within a sentence. As such, we hypothesize that other methods, such as Brown clusters [19] or topic model labels, would suffice. In the case where no word clustering method at all is available nor can be induced for the language, it seems doubtful that one could have enough data where data selection would do any good.

6. Conclusion

The data selection method of [1] directly uses the fact that the in-domain and general corpora differ in order to quantify the relevance of sentences in a data pool to an in-domain task text. This relevance is based on how much a sentence is like the in-domain corpus and unlike the pool corpus.

We have presented a way to further leverage the discriminative mechanics of the Moore-Lewis data selection process to distill a corpus down to a representation that explicitly encodes differences between the corpora for the specific data selection scenario at hand. We do this by replacing every word in the corpora with its part-of-speech tag plus a suffix that indicates the relative bias of the word, or how much likelier it is to be in the task corpus versus the pool.

Language models trained on data selected with our approach have -13 lower absolute perplexity on in-domain data than the baselines, doubling the effectiveness of the cross-entropy difference based method. The trained language models also had 37% fewer OOV's on the task data than the standard baseline. Furthermore, machine translation systems trained on data selected with our approach outperform MT systems trained on data selected with regular n-gram models by up to +1.8 BLEU, or can be stacked with in-domain translation model for up to +1.3 BLEU. These improvements come despite using zero of the original words in the texts for our selection process, and reducing the corpus vocabulary to under 200 automatically-derived tags.

By changing the representation of the text, we can use basic n-gram models to characterize the difference between the corpora. This process enables us to use common models with robust statistics that are tailored to computing the similarity score, instead of training a separate classifier or ignoring the textual differences as the standard approach does.

As a bonus, our new representation and language difference models mean that the data selection process itself is now no longer memory-bound. Because the corpus vocabulary is so compact, the language models required are also much smaller, and ordinary computational resources now suffice to perform data selection on practically any size corpus.

Much work remains, as there are surely other useful factors and more nuanced representations. What else is there about a task that differentiates its language from others, how can we quantify these features, and which of them are useful when measuring the difference between two texts? We have not explored the parameter space for our approach, either. One might wish in the future to try use powers of 2, or e, or linear bucket ranges, or adjust the ranges to ensure words are evenly distributed amongst buckets. Furthermore, one might not want to collapse the most discriminative words – the ones with the highest contribution to the cross-entropy difference score – into the same classes based on POS tag. It might be the case that it is only important to lump the least discriminative words together so as to focus the selection model on the differences between the corpora.

7. Acknowledgements

We appreciate the helpful feedback of Philip Resnik and the anonymous reviewers.

8. References

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

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

[3] S. Mansour, J. Wuebker, and H. Ney, "Combining Translation and Language Model Scoring for Domain-Specific Data Filtering," *IWSLT (International Workshop on Spoken Language Translation)*, 2011.

[4] L. Wang, D. F. Wong, L. Chao, J. Xing, Y. Lu, and I. Trancoso, "Edit Distance : A New Data Selection Criterion for Domain Adaptation in SMT," *RANLP (Recent Advances in Natural Language Processing)*, no. September, pp. 727–732, 2013.

[5] K. Duh, G. Neubig, K. Sudoh, and H. Tsukada, "Adaptation Data Selection using Neural Language Models: Experiments in Machine Translation," *ACL (Association for Computational Linguistics)*, 2013.

[6] H. Cuong and K. Sima'an, "Latent Domain Translation Models in Mix-of-Domains Haystack," *COLING (International Conference on Computational Linguistics)*, 2014.

[7] A. Axelrod, P. Resnik, X. He, and M. Ostendorf, "Data Selection With Fewer Words," *WMT (Workshop on Statistical Machine Translation)*, 2015.

[8] J. Burrows, "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing*, vol. 17, no. 3, pp. 267 –287, 2002.

[9] D. Biber, *Variations Across Speech and Writing*. Cambridge, UK: Cambridge University Press, 1988.

[10] H. Baayen, H. V. Halteren, and F. Tweedie, "Outside the cave of shadows: using syntactic annotation to enhance authorship attribution," *Literary and Linguistic Computing*, vol. 11, no. 3, pp. 121–132, 1996.

[11] S. Argamon, M. Koppel, and G. Avneri, "Routing documents according to style," *Workshop on Innovative Information Systems*, vol. 60, no. 6, pp. 581–3, 1998.

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

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

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

[15] C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blumson, H. Setiawan, V. Eidelman, and P. Resnik, "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*, 2010.

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

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

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

[19] P. F. Brown, P. V. DeSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-Based n-gram Models of Natural Language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.