Exploiting Relative Frequencies for Data Selection

Thierry Etchegoyhen tetchegoyhen@vicomtech.org
Andoni Azpeitia aazpeitia@vicomtech.org
Eva Martinez García emartinez@vicomtech.org
Vicomtech-IK4, Donostia / San Sebastián, Gipuzkoa, Spain

Abstract

We describe a data selection method for domain adaptation in machine translation, based on relative frequency ratios computed between in-domain and out-of-domain corpora. Our method is compared to a state-of-the-art approach based on cross-entropy differences, outperforming it significantly in terms of data sparseness reduction and BLEU scores on the models created from various data slices. This approach is also shown to either perform significantly better or provide competitive results in terms of perplexity when compared to a method designed to minimise cross-entropy. A novel method to mine unknown words in out-of-domain datasets is also presented, resulting in the best models across the board when used to weight sentences whose similarity to the primary domain is determined by relative frequency ratios. The proposed method is simple, requiring neither external resources nor complex setups, which makes it highly portable across domain adaptation scenarios.

1 Introduction

Data-driven approaches to machine translation, such as statistical machine translation (SMT) (Brown et al., 1990) or neural machine translation (NMT) (Bahdanau et al., 2014), need large volumes of quality bilingual data to be trained effectively. In most scenarios, machine translation systems trained only on available in-domain bilingual corpora face data sparseness issues which hinder on their coverage and accuracy. Identifying useful subsets of out-domain data through automated bilingual data selection has thus become an important method for domain adaptation.

While no fully accurate method has been designed yet to identify subsets of out-of-domain data that are useful and sufficient to improve machine translations models, the main characteristics of the data being sought can be safely assumed to cover two main aspects.

First, the selected out-of-domain sentences should cover lexical and syntactic gaps in the domain, in order to improve the in-domain translation models. This aspect can be controled by measuring the amount of unknown words after incorporating the selected data and by evaluating the impact of the out-of-domain data on the quality of the resulting machine translation models via automated metrics.

Secondly, the selected data should not add confusion to the models, an aspect which can be measured in terms of language model perplexity on either side of the bilingual data. A method that would partially cover lexical and syntactic gaps while also adding significant subsets of data unrelated to the domain at hand would be adding statistical noise to the translation models, thus lowering their accuracy.

These two aspects in combination render the selection task particularly difficult, as the optimal target data should be similar to the in-domain data, in order to reduce the noise added to the models, while also provide enough new material to improve the primary models.

In this paper, we explore the potential of a simple approach based on relative frequency ratios between in-domain and out-of-domain distributions. We evaluate the benefits of our approach in two domain adaptation scenarios featuring large volumes of out-of-domain data and compare it to a state-of-the-art data selection method based on bilingual cross-entropy differences. We show that our approach outperforms data selection based on cross-entropy, achieving significantly better results in terms of translation metrics, while also significantly reducing the amount of out-of-vocabulary words and equating or improving perplexity results.

The remainder of this article is organised as follows: Section 2 summarises related work in bilingual data selection, in particular for domain adaptation; Section 3 describes the proposed approach based on relative frequencies; in Section 4, we present the corpora, models and results of our comparative experiments; finally, in Section 5 we draw conclusions from this work.

2 Related Work

Selecting subsets of bilingual corpora has been a popular approach to create domain-adapted and/or more compact translation systems, see (Eetemadi et al., 2015) for a recent detailed survey. For domain adaptation in particular, the main goal has been to better exploit available parallel corpora, by selecting the minimal subsets of bilingual sentence pairs that maximise the accuracy gains of machine translation systems for a specific domain, where training resources are usually scarce.

Several approaches have been explored over the years for bilingual data selection. TF-IDF weighting has been used for instance by (Lü et al., 2007) for similar sentence identification and weighted training, and by (Eck et al., 2005), who combine it with unseen n-gram frequency scoring to create competitive SMT systems based on smaller training sets. Foster et al. (2010) first rank the out-of-domain sentence pairs according to the perplexity of the in-domain target side language model, then retain the number of top-ranked pair that maximizes the BLEU score on a development set. They further refine the selection process by extending the weight learning approach in Matsoukas et al. (2009), through phrase pair weighting, feature-based measures of the usefulness of phrases and incorporating instance-weighting into a linear combination model.

Perplexity-based methods have figured prominently in work focusing on bilingual data selection. (Foster et al., 2010), for instance, use in-domain target side perplexity to rank out-of-domain sentence pairs and select top-ranked pair that maximize the BLEU score on held-out sets, whereas (Mansour et al., 2011) combine language model and translation model cross-entropy scores to the task of data selection. In (Aydın and Ozgür, 2014), the out-of-domain corpora are ranked according to in-domain perplexity and proper subsets of the data are selected using the vocabulary saturation technique of (Lewis and Eetemadi, 2013). One the most popular data selection methods is that of (Axelrod et al., 2011), who extend work by (Moore and Lewis, 2010) by ranking out-of-domain sentences according to bilingual cross-entropy differences as determined by source and target in-domain and out-of-domain language models.[1] Cross-entropy differences select sentences that are both similar to the in-domain data, and unlike the average out-of-domain data. Generalisations of word-based cross-entropy differences have been proposed by (Axelrod et al., 2015a) and (Axelrod et al., 2015b), improving over the standard approach by means of part-of-speech generalisation and class-based language models.

Whereas most approaches attempt to design optimal similarity measures between domains, (Banerjee et al., 2012) use translation quality to guide data selection. In their approach, batches of out-of-domain data are incrementally added to an existing baseline system, evaluated in terms of translation quality on a development set, and a given batch is selected only if its inclusion improves translation quality. Also not focused on sentence similarity is work by (Daumé III and

[1] We will refer to their approach as Modified Moore-Lewis (MML), although it has received a large variety of acronyms in the literature.

Jagarlamudi, 2011), who address the lexical coverage aspect of supplementary data selection by mining unknown words via canonical correlation analysis. (Gascó et al., 2012) use approximations of in-domain probability distributions and n-gram infrequency scores to achieve significant improvements over the baselines and over random selection. In recent work, (Wong et al., 2016) report significant improvements over perplexity-based selection for Chinese-English, by training recurrent neural networks to select supplementary data.

Selection based on bilingual cross-entropy differences can be considered the *de facto* state-of-the-art approach and is standardly used as baseline by competing approaches. In (Kirchhoff and Bilmes, 2014), the use of submodular functions for data selection obtained minor but statistically significant BLEU score gains over MML, whereas (Peris et al., 2017) achieve slight improvements in terms of BLEU scores via neural network-based classification while using less data. (Banerjee et al., 2013) also compare their data selection method, based on quality estimation, to MML and obtain slightly better BLEU scores while using smaller amounts of data as well. Overall, albeit statistically significant in most reported cases, improvements over MML have been small in terms of automated translation metrics and this method can thus still be considered a strong baseline for comparative evaluations.

Although we focus our work on bilingual data selection, it is worth noting that monolingual data selection for language model adaptation has also been a fruitful approach, explored in several studies. (Mediani et al., 2014), for instance, improve over cross-entropy selection by drawing better samples of out-of-domain data and using word association as a mean to add semantic similarity into the selection process. (Mansour et al., 2011) describe a filtering approach based on combined cross-entropy scores for the language and translation models, and report small but statistically significant improvements over standalone methods. Recently, (Duh et al., 2013) have explored the use of neural language models for data selection, and in particular the advantages of continuous vector spaces over n-gram-based approaches on handling unknown words in out-of-domain corpora. We leave monolingual data selection aside in what follows, although we believe our approach to be worth exploring on these grounds as well, given the results in terms of perplexity and data sparseness reduction described in Section 4.

3 Exploiting Relative Frequencies

By their very nature, perplexity-based approaches tend to favour short out-of-domain sentences that exhibit n-gram distributions close to the primary domain. Although this has been shown to be a fruitful approach in some data selection scenarios, it leaves aside the potential contribution of data that is related to the primary domain while also exhibiting different distributions. In the worst case, perplexity-based methods could select out-of-domain sentences that are already present in the in-domain pool, thus defeating the purpose of increasing model coverage using additional data. Although this is not usually the case with contrasted domains, the main expectation is that machine translation models should benefit more from additional data that cover both lexical gaps and unseen syntactic configurations. In order to test this hypothesis, we design a method that scores out-of-domain sentences according to their similarity to the domain of interest, while not biasing selection towards the in-domain n-gram distributions.

3.1 Relative Frequency Ratios

The approach we propose estimates similarity via relative frequency ratios between the in-domain and out-of-domain data. More specifically, we first compute relative frequencies for each word w in corpus c through token counts C as in Equation 1:

$$\phi_c(w) = \frac{C(w)}{\sum_{i=1}^{|c|} C(w_i)} \tag{1}$$

For each out-of-domain pair (s, t), where s is the set of source words and t the set of target words, the relative frequency score is then computed as the sum of the ratios of in-domain and out-of-domain relative frequencies as in Equation 2, taking the arithmetic mean of the scores for the source and target sentences.

$$rfr(s,t) = \frac{\sum_{i=1}^{|s|} \frac{\phi_d(w_i)}{\phi_o(w_i)} + \sum_{i=1}^{|t|} \frac{\phi_d(w_i)}{\phi_o(w_i)}}{2} \quad (2)$$

In the above equation, ϕ_d and ϕ_o denote the relative frequencies computed on the in-domain and out-of-domain corpora, respectively. To reduce the impact of large differences in terms of sentence length, scoring is applied to the sets of tokens composing each sentence. Out-of-domain words that are not represented in the in-domain corpus are ignored, as the frequency ratio would not be computable in this case.

The metric thus favours sentences with the largest amount of words that are more represented in the in-domain than in the out-of-domain. Additionally, it refrains from ignoring the relative distributions of frequent words, such as function words, under the assumption that all words in a given sentence are important to identify similarity as defined in terms of content, register and style. Finally, the metric remains neutral regarding out-of-in-domain words, as they do not impact the score of a given sentence, and does not favour known n-gram distributions as it is based solely on cumulative word frequency ratios.[2]

3.2 Mining Unknown Words

As previously mentioned, the metric described in the previous section ignores words that are not part of the in-domain vocabulary. Out-of-domain sentences that contain out-of-vocabulary (OOV) words will thus be scored according only to the words in the sentence that do pertain to the in-domain data. This might not be optimal in two respects.

First, large pools of out-of-domain corpora are typically noisy, containing, for instance, sentences in languages other than the expected ones or sequences of corrupt characters. The similarity score in this case would only be determined by the known words, typically punctuation, leaving aside the fact that sentences containing mostly OOV words are not likely to improve the translation models.

Secondly, since adding corpora to the models aims at improving model coverage on both lexical and syntactic grounds, sentences that resemble the in-domain data while also providing new vocabulary, and by extension, phrases, should be favoured over those that are only based on known vocabulary.

Taking these two aspects into account, we aim to promote those sentences that exhibit a reasonable amount of out-of-vocabulary words and minimise the score of those with large amounts of OOV words. In order to do so, we complement the core metric with a weighting scheme related to the percentage of OOV words in each out-of-domain sentence.

Let u be the percentage of out-of-vocabulary items in an out-of-domain sentence, given the in-domain vocabulary. The value u is first assigned a weight according to Equation 3:

$$W(u) = sin(\alpha \cdot u^k) \quad (3)$$

Using this sinusoidal function over percentage of unknown words gives us the expected behaviour: sentences above a given threshold of OOV words will be scored negatively, while the

[2] In additional experiments not reported here, data selection based on relative frequency ratios performed better than log-likelihood-based termhood as in (Gelbukh et al., 2010). We hypothesize that this result is due to the latter both ignoring words that have higher relative frequency in the out-of-domain corpus and to the relative demotion of weaker terms.

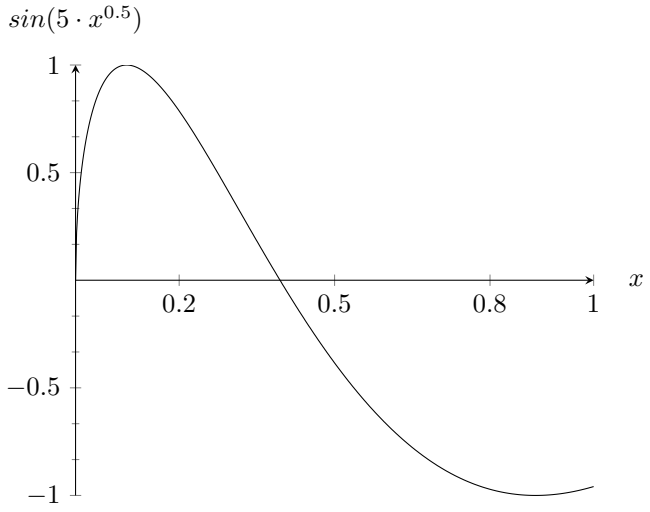

Figure 1: Graph of weighting function for the selected hyper-parameters

higher values from this function will be obtained with a small percentage of unknown words; sentences containing only known words are assigned a value of zero.

The hyper-parameters α and k need to be set empirically, according to how aggressively one wants to mine out-of-vocabulary items. The graph of the function for hyper-parameters α=5 and k=0.5, which were the ones selected for the experiments reported here, is given in Figure 1.

With the selected hyper-parameters, sentences with a percentage of unknown words around 10% will thus be promoted while amounts over 40% will be considered detrimental. The integration of this weighting scheme to the core metric is described in Equation 4.

$$wrfr(s,t) = \frac{exp(W(u_s)) \cdot \sum_{i=1}^{|s|} \frac{\phi_d(w_i)}{\phi_o(w_i)} + exp(W(u_t)) \cdot \sum_{i=1}^{|t|} \frac{\phi_d(w_i)}{\phi_o(w_i)}}{2} \quad (4)$$

Since the core metric is based on relative frequency sums while the weighting scheme ranges over positives and negatives, the values of the W function are mapped to the positive space via exponentiation. Thus, sentences with no unknown words are scored only according to their relative frequency ratios, those with amounts above forty percent will receive a weight between 0 and 1, and the remainder will be favoured with weights above 1. In the next sections, we evaluate the impact of this weighting scheme on the data selection process.

4 Experiments

The experiments described in this section were designed to compare data selection methods in realistic scenarios, where only a fraction of the large out-of-domain data is typically sought. The out-of-domain data, as ranked by each method, were thus sliced from one percent up to twenty percent of the data to perform the evaluations reported here. We compare the two variants of our approach to Modified Moore-Lewis as representative of the state of the art among methods that do not require sophisticated setups and are thus easily portable across domain adaptation scenarios.[3]

[3] MML data selection was performed with the XenC tool (Rousseau, 2013): https://github.com/rousseau-lium/XenC.

LANGS	CORPUS	TRAIN	DEV	TEST
EN-ES	NewsCommentary	207,137	3,003	3,000
EN-FR	EMEA	354,288	500	1000

Table 1: In-domain corpora

LANGS	CommonCrawl	Europarl	UN	POOL
EN-ES	1,814,883	1,842,496	8,079,790	11,661,326
EN-FR	3,065,194	1,826,770	9,142,161	13,864,506

Table 2: Out-of-domain corpora

The data slices used here are similar to those employed by (Axelrod et al., 2011), who experimented with subsets corresponding to 1 time, 2 times and 4 times the size of the in-domain corpus, and by (Axelrod et al., 2012), who opted to select 10% of the out-of-domain data for all of their experiments.[4]

4.1 Corpora

As in-domain corpora, for English-Spanish we used the *NewsCommentary* datasets from the WMT news translation shared task, with *newstest2012* as development set and *newstest2013* as test set; for English-French we used the data from the WMT medical translation task, with EMEA as training set and the *khresmoi-summary* development and test sets.[5]

As out-of-domain corpora, we used three of the available corpora in the aforementioned WMT tasks, namely: *CommonCrawl*, *Europarl* and *UN*. All three corpora were pooled in a single corpus, whose data was then ranked by each method. The statistics for the corpora, after filtering sentences larger than 60 tokens and removing duplicates, are shown in Table 1 and Table 2.

This setup responds to two main goals. First, data selection is applied to a large pool of publicly available out-of-domain data composed of three different sub-domains with varying amounts of noise.[6] This allows for an evaluation of the robustness of each method.

Secondly, the in-domain datasets were selected to be largely different, one covering news commentary and the other medical data, while the out-of-domain data pool remains constant. This provides results in a data selection scenario where the out-of-domain datasets are not pre-selected according to their closeness to the in-domain data. It also allows for a contrastive evaluation of the benefits of the same out-of-domain data for different in-domain corpora.

4.2 Selected Data

The compared methods select different subsets of data at every slice, as shown in Table 3. This is not unexpected given their respective scoring schemes, with short pseudo in-domain sentences being targeted by MML and longer lexically similar sentences by cumulative frequency ratios. The two variants of our approach have a significant amount of common selected sentences on the EN-FR dataset, but more than half of the sentences they select are different up to the 10% slice in EN-ES. This demonstrates the marked impact of the technique we adopted to mine unknown words on the selection of sentences deemed similar by their relative frequency ratios.

The results in Table 3 also show that the two adaptation scenarios differ markedly with

[4]Note that the dichotomic search for optimal slices, computed by the XenC tool using perplexity scores on held-out sets, identified best points at 1% and 14% for the English-French and English-Spanish datasets, respectively.

[5]All datasets are available at http://www.statmt.org/wmt13/ and http://www.statmt.org/wmt14/.

[6]The *CommonCrawl* corpus contains large sections of noisy data, for instance.

LANGS	METHODS	1PCT	2PCT	5PCT	10PCT	20PCT	30PCT	40PCT	50PCT
EN-ES	MML-RFR	11.72	15.88	23.86	32.59	44.12	52.88	60.59	67.91
	MML-WRFR	5.16	7.19	12.17	19.12	30.85	41.39	51.30	60.79
	RFR-WRFR	44.81	43.72	45.42	49.65	57.51	64.40	70.81	77.13
EN-FR	MML-RFR	24.55	24.30	24.72	27.31	34.45	42.08	49.90	57.95
	MML-WRFR	21.09	21.30	22.08	24.88	32.13	39.94	48.04	56.49
	RFR-WRFR	84.20	83.66	82.59	82.59	83.70	85.04	86.39	87.81

Table 3: Percentage of common selected sentence pairs per data slice

LANGS	METHOD	1PCT	2PCT	5PCT	10PCT	20PCT	30PCT	40PCT	50PCT
EN-ES	MML	17.5	19.3	21.6	23.2	24.6	25.4	25.8	26.1
	RFR	40.11	40.03	39.18	37.90	36.03	34.63	33.44	32.30
	WRFR	39.41	40.04	39.85	38.86	37.07	35.52	34.08	32.70
EN-FR	MML	14.2	15.7	17.9	19.7	21.8	23.0	23.7	24.2
	RFR	28.44	29.72	31.86	33.14	33.56	33.10	32.31	31.35
	WRFR	29.77	31.04	33.12	34.19	34.29	33.60	32.63	31.53

Table 4: Average source sentence length per data slice

respect to the distribution of similar sentences in the out-of-domain corpus. Whereas there is a significant portion of similar material in the in-domain and out-of-domain data for EN-ES, due to the presence of the *Europarl* and *UN* corpora, for EN-FR the amount of out-of-domain data related to the medical domain is sparser. This produces the expected larger selection differences between methods in the first case as compared to the second one.

As previously noted, the methods are expected to differ in terms of length, given their respective scoring schemes. The comparative data shown in Table 4 confirm this expectation, with the two methods based on relative frequency ratios selecting sentences that are on average double the length of those selected by MML in the first three slices. Length differences tend to reduce in larger slices, although the perplexity-based approach still tends to select shorter sentences overall.

All three methods select data that seem related to the in-domain at first glance, as illustrated in Table 5 with examples of the type of out-of-domain sentences uniquely selected by each method in their respective 1% slices. In the next sections, we measure more precisely how related the selected data are to the in-domain in terms of capturing out-of-vocabulary words, perplexity and automated translation metrics.

4.3 Unknown Words

One of the main motivations for an approach based on cumulative frequency ratios is its selection of longer sentences similar to the in-domain, which is meant to increase the amount of unknown words that can be captured, indirectly in the case of RFR and directly in the case of WRFR.

To evaluate the differences between the three methods in terms of increasing in-domain coverage, we measured the number of out-of-vocabulary items a posteriori on the test sets for each data slice. Figure 2 shows the results on the source side in EN-ES.

For this language pair, the amount of OOV items under MML is more than double that of WRFR, and nearly double that of RFR, for the lower slices. At the 1% mark in EN-ES, for instance, the slices contain 2669, 1529 and 1146 unknown words when selected by MML, RFR and WRFR, respectively. The amounts of OOV words only start to be similar around the 50

LANGS	METHOD	SENTENCES
EN-ES	MML	where are we heading ? — trillions of dollars more are waiting in the wings . — the implications are dire .
	RFR	the assumption that only an enlightened minority is in a position to respect human rights and freedoms . — greenhouse gas emissions can be cut through the use of nuclear energy , clean coal and low carbon-emitting renewable energies . — coupled with extensive deregulation of financial markets and excess liquidity , these imbalances encouraged investors to engage in leveraged risk-taking in search of profits .
	WRFR	during that period , their debt actually increased from $ 618 billion in 1980 to $ 3.25 trillion in 2006 . — Mr. Snowden (United States of America) said that the Commission for Sustainable Development had galvanized action and helped shape the agendas of a wide range of organizations around the world . — there has been a temptation for the West – Europe and the United States – to stress continuity and so-called stability .
EN-FR	MML	avoid contact with skin , eyes or clothing . — the unused portion should be discarded . — peel open the package with dry hands and place the tablet on your tongue .
	RFR	in terms of public health , the environmental impact of the new medicinal products should be assessed . — antiretroviral treatment can be effective only if it is administered and monitored by health professionals working in a well-functioning national health system . — finally , it recognises the need for studies on vaccines and anti-viral medications that are independent of the pharmaceutical industry , including with regard to the monitoring of vaccination coverage .
	WRFR	during the final process , an operator peers through a microscope at the die surfaces , polishing them carefully with a diamond abrasive tool head that is vibrated by supersonic waves . — concentrations of petroleum contaminants in fish and crab tissue , as well as contamination of shellfish could have potentially significant adverse effects on health . — the first three , namely glycerine , brake fluid and anti-freeze , are considered to present the most extreme incompatibility with calcium hypochlorite .

Table 5: Uniquely selected English sentences in 1% slices

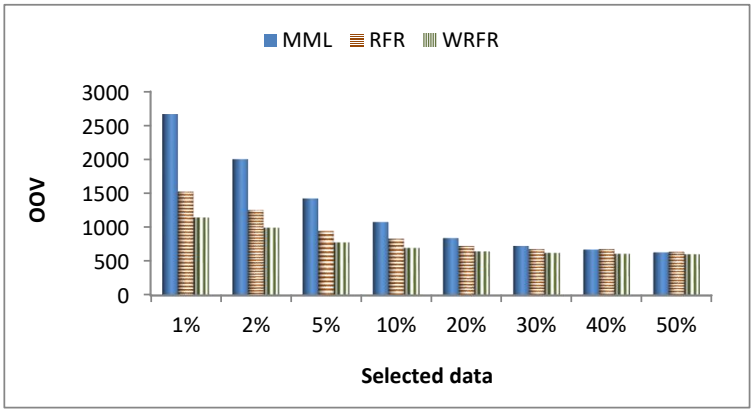

Figure 2: Source out-of-vocabulary words per English-Spanish data slice

percent mark, although WRFR still captures more unknown words in all cases.

The results for the source side in EN-FR are shown in Figure 3.

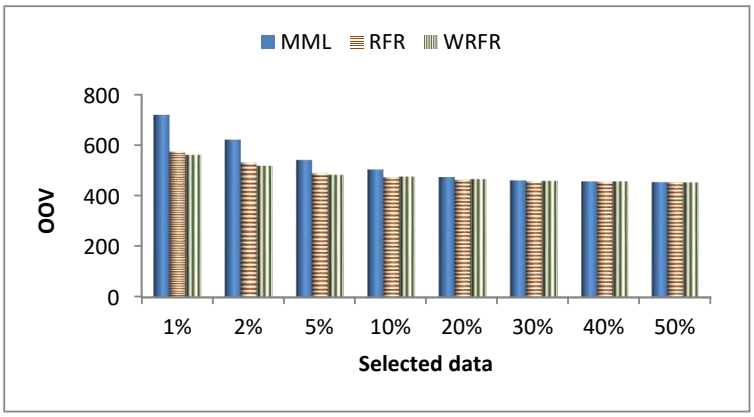

Figure 3: Source out-of-vocabulary words per English-French data slice

For this language pair, the tendencies are similar, with MML being markedly outperformed by both RFR and WRFR, and the latter being the best of all three methods in terms of selecting out-of-domain data that reduce data sparseness.

4.4 Perplexity

As seen in the previous section, the compared approaches differ significantly in terms of selected data, in terms of both average length and amount of OOV items they capture. In addition to these measures, it is important to evaluate the increase or reduction in their respective statistical modelling of sequences. Table 6 indicates the perplexities, including OOV words in the computation of entropy, obtained by each method on the respective test sets after training language models on each data slice.

Since MML is designed to target those sentences that have low in-domain perplexity and high out-of-domain perplexity, one could expect this method to significantly outperform methods based on relative frequency ratios, which make no attempt at minimizing perplexity. As shown above, this is not the case, with both RFR and WRFR significantly outperforming MML

LANG	METHOD	1PCT	2PCT	5PCT	10PCT	20PCT	30PCT	40PCT	50PCT
ES	MML	335.55	295.46	249.95	217.95	**196.59**	**188.95**	**186.60**	**186.60**
	RFR	281.53	252.06	224.52	210.23	201.26	198.49	197.38	197.17
	WRFR	**257.67**	**232.76**	**211.72**	**202.32**	197.93	197.15	196.85	196.79
FR	MML	**151.90**	**147.55**	**151.63**	**161.38**	**175.67**	**187.42**	**196.67**	**203.95**
	RFR	153.63	154.66	163.54	173.54	187.74	197.88	205.13	210.47
	WRFR	157.64	158.75	166.89	177.64	191.15	200.44	207.20	212.14

Table 6: Target language perplexity with OOV per data slice

up to the 10% slice in EN-ES. For EN-FR, the MML approach provides better results on all slices, but only marginally so when compared to the differences obtained in EN-ES.

With both RFR and WRFR outperforming MML in terms of OOV coverage, it could be hypothesised that the competitive results in terms of perplexity are largely due to differences in the amount of captured unknown words. To evaluate this specific point, we computed perplexity using the same language models but ignoring OOV words, with the results shown in Table 7.

LANG	METHOD	1PCT	2PCT	5PCT	10PCT	20PCT	30PCT	40PCT	50PCT
ES	MML	221.38	211.69	193.56	177.30	**165.77**	**162.42**	**161.72**	**161.72**
	RFR	217.07	202.41	186.75	178.65	173.77	172.45	172.28	172.57
	WRFR	**211.53**	**196.18**	**182.75**	**176.67**	173.97	173.61	173.51	173.38
FR	MML	**116.81**	**116.01**	**121.36**	**129.67**	**140.90**	**149.75**	**157.47**	**163.06**
	RFR	123.81	125.72	132.94	140.82	151.35	159.15	164.47	168.46
	WRFR	127.86	128.82	135.70	144.10	154.29	161.20	166.20	169.95

Table 7: Target language perplexity without OOV per data slice

The tendencies observed for perplexity including all words are maintained, with RFR and WRFR obtaining the best scores on the first slices in EN-ES and MML obtaining lower perplexities across the board for EN-FR. The differences between the methods are less marked in this case, due to ignoring out-of-vocabulary items in the computation of perplexity.

Overall, the two variants based on relative frequencies perform well in terms of perplexity, either outperforming the perplexity-minimising MML approach or reaching comparable results.

4.5 Extrinsic Evaluation

Finally, we performed extrinsic evaluations using SMT models trained on the in-domain and out-of-domain corpora, as there exist well-established methods to perform domain adaptation with said models. All translation models are phrase-based (Koehn et al., 2003), trained using the Moses toolkit (Koehn et al., 2007) with default hyper-parameters and phrases of maximum length 5. The phrase tables were pruned according to statistical significance (Johnson et al., 2007) and the parameters of the log-linear models were tuned with MERT (Och, 2003). All language models are of order 5, trained with the KENLM toolkit (Heafield, 2011). The individual in-domain and out-of-domain translation models were then combined by filling up the in-domain phrase table with out-of-domain phrases, with a binary feature denoting the origin of each phrase (Bisazza et al., 2011).

We did not perform additional extrinsic evaluations using neural machine translation models for this work. Although this could provide valuable additional information, domain adaptation with NMT is an ongoing research activity where current approaches have certain limitations. One of the main methods currently employed is that of specialisation, where a network trained

MODEL	100PCT	1PCT	2PCT	5PCT	10PCT	20PCT
NEWSCOM	23.285	_	_	_	_	_
POOL	27.746	_	_	_	_	_
RAND	_	24.065	24.246	25.194	26.273	27.01
MML	_	23.637	24.121	24.999	26.101	26.878
RFR	_	24.547 † ‡	25.102 † ‡	26.0 † ‡	26.563 † ‡	27.166 ‡
WRFR	_	**24.823** † ‡ ∗	**25.458** † ‡ ∗	**26.142** † ‡	**26.914** † ‡ ∗	**27.258** † ‡

Table 8: BLEU scores per data slice for English-Spanish

MODEL	100PCT	1PCT	2PCT	5PCT	10PCT	20PCT
EMEA	27.099	_	_	_	_	_
POOL	37.958	_	_	_	_	_
RAND	_	31.564	31.747	33.234	34.52	37.43
MML	_	33.695 †	34.805 †	35.907 †	36.539 †	37.43
RFR	_	34.791 † ‡	**35.48** † ‡	**35.979** †	**37.438** † ‡ ∗	37.276
WRFR	_	**35.124** † ‡ ∗	**35.325** † ‡	**36.298** †	36.987 † ‡	37.268

Table 9: BLEU scores per data slice for English-French

on generic data is subsequently extended with additional in-domain data (Crego et al., 2016). As it stands, this method requires the new data to be constrained to the vocabulary of the already trained network, which prevents a direct contribution of in-domain vocabulary. This specific issue is typically mitigated via external dictionaries along with a copy mechanism for words that are not part of the generic vocabulary, a working solution which does not however allow for a complete modelling of the in-domain data. Adopting a reversed approach would result in training an in-domain model and add the selected out-of-domain data, as is typically done in SMT domain adaptation. However, the networks would specialise towards the selected out-of-domain data in this case, which would not provide the expected domain adaptation results. A third approach would be to train several models from scratch using a combination of all the in-domain data along with each slice of selected out-of-domain data, a highly computationally expensive approach since each addition of selected data slices would require the training of an entire network.

The time-tested SMT-based approach we chose for our experiments has the advantage of not putting internal restrictions on the available vocabulary and allows for a straightforward comparison between the different data selection approaches. We thus leave additional NMT-based contrastive experiments for future work, noting that evaluating the contribution of selected portions of out-of-domain data, as determined by each one of the compared methods, on NMT models, would undoubtedly provide interesting additional results.

In addition to the models trained on each slice as selected by the three compared methods, for both scenarios we trained a POOL model by combining the in-domain model with a model trained on all out-of-domain data, and used randomly sampled data to train a random baseline (RAND). The comparative results for the two domain adaptation scenarios in terms of BLEU scores (Papineni et al., 2002) are shown in Tables 8 and 9 for English-Spanish and English-French, respectively.[7]

[7]Statistical significance was measured using the paired bootstrap resampling test of (Koehn, 2004) over average BLEU scores. † indicates statistical significance, at $p < 0.05$, as computed between a given model and the random baseline; ‡ between RFR or WRFR and MML; and ∗ between RFR and WRFR.

Overall, the RFR and WRFR approach outperformed MML across the board, with results exhibiting no statistically significant differences between the three models obtained only at the 20% slice mark in EN-FR. Using only 1% of the data, the WRFR approach improves over MML by 1.2 BLEU points for EN-ES and by 1.4 for EN-FR. Given the usually minor improvements obtained by alternatives against MML, these results are indicative of the ability of approaches based on relative frequencies to select useful data that help reach significant improvements of machine translation models.

Note that, for EN-ES, there is no statistically significant difference between MML and the random baseline, although all methods perform better than random selection throughout in EN-FR. This difference might be attributable to the fact that MML tends to select already known material, which is more likely to be selected in this out-of-domain pool when the in-domain contains news-related data. Thus, the selected data bring less new and useful data than is the case for EN-FR, where there is a wider gap between medical in-domain data and the average data in the out-of-domain pool.[8]

Also interesting to note is the fact that no model performs better than the ones created with all out-of-domain data. Note that several reports of models trained on a subset of the data having outperformed the reference models trained on all data indicated such results when using larger slices than the ones reported here (see e.g., Banerjee et al. (2012); Wong et al. (2016)); in other cases, the best results do not outperform the larger models (see, e.g., Peris et al. (2017)). Results on these grounds are also largely dependent on the volumes and nature of out-of-domain data being used; in our case, the pools are on the larger side of the reported experimental scales and contain merged data from different domains, which renders the task more difficult for any data selection method. In any case, the methods evaluated here already reach results that are close to those obtained using all the available data, while using only a fraction of the data, which is one of the main reasons to apply data selection.

5 Conclusion

We described a data selection method for domain adaptation in machine translation, based on relative frequency ratios computed between in-domain and out-of-domain corpora. Our method was compared to a state-of-the-art approach based on cross-entropy differences, outperforming it in terms of data sparseness reduction and BLEU scores on the models created from various data slices. Although not meant to minimise perplexity, our approach was shown to either perform significantly better with fewer data or provide competitive results.

A novel method to mine unknown words in out-of-domain datasets was also presented, which resulted in the best models across the board when used to weight sentences whose similarity to the primary domain was determined by relative frequency ratios. This empirical method can be applied to other scenarios as well, where the goal is to target sentences according to the desired amount of unknown words.

The proposed method is simple, requiring neither external resources nor complex setups, which makes it highly portable across domain adaptation scenarios. In future work, we will pursue improvements and comparative evaluations of the presented methods, in particular with neural machine translation models, where the comparatively larger amounts of useful data retrieved by the method we described might also contribute to increase model accuracy.

[8] Since MML depends on sampling the out-of-domain data in similar proportion to the size of the in-domain, different samples might give different results, especially on large datasets. Several samples could be drawn from the same out-of-domain datasets, a functionality that is provided by the XenC toolkit. However, results along these lines have not been fully explored, to the best of our knowledge, and we opted to use the MML method in its standard variant with single sampling.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics.

Axelrod, A., He, X., Resnik, P., and Ostendorf, M. (2015a). Data selection with fewer words. pages 58–65.

Axelrod, A., Li, Q., and Lewis, W. D. (2012). Applications of data selection via cross-entropy difference for real-world statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 201–208.

Axelrod, A., Vyas, Y., Martindale, M., Carpuat, M., and Hopkins, J. (2015b). Class-based n-gram language difference models for data selection. In *Proceedings of the 12th International Workshop on Spoken Language Translation*.

Aydın, B. and Ozgür, A. (2014). Expanding machine translation training data with an out-of-domain corpus using language modeling based vocabulary saturation. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA)*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.

Banerjee, P., Naskar, S. K., Roturier, J., Way, A., and van Genabith, J. (2012). Translation quality-based supplementary data selection by incremental update of translation models. In *Proceedings of COLING 2012: Technical Papers*, pages 149–166.

Banerjee, P., Rubino, R., Roturier, J., and van Genabith, J. (2013). Quality estimation-guided data selection for domain adaptation of smt. *MT Summit XIV: proceedings of the fourteenth Machine Translation Summit*, pages 101–108.

Bisazza, A., Ruiz, N., Federico, M., and Kessler, F.-F. B. (2011). Fill-up versus interpolation methods for phrase-based smt adaptation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 136–143.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., et al. (2016). Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Daumé III, H. and Jagarlamudi, J. (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*, pages 407–412. Association for Computational Linguistics.

Duh, K., Neubig, G., Sudoh, K., and Tsukada, H. (2013). Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 678–683.

Eck, M., Vogel, S., and Waibel, A. (2005). Low cost portability for statistical machine translation based on n-gram coverage. In *Proceedings of MTSummit X*.

Eetemadi, S., Lewis, W., Toutanova, K., and Radha, H. (2015). Survey of data-selection methods in statistical machine translation. *Machine Translation*, 29(3-4):189–223.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 451–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 152–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gelbukh, A., Sidorov, G., Lavin-Villa, E., and Chanona-Hernandez, L. (2010). Automatic term extraction using log-likelihood based comparison with general reference corpus. In *International Conference on Application of Natural Language to Information Systems*, pages 248–255. Springer.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johnson, J. H., Martin, J., Foster, G., and Kuhn., R. (2007). Improving translation quality by discarding most of the phrasetable. In *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 967–975.

Kirchhoff, K. and Bilmes, J. (2014). Submodularity for data selection in statistical machine translation. In *Proceedings of the 2014 Conference on Empricial Methods in Natural Language Processing*.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics-Volume 1*, pages 48–54. Association for Computational Linguistics.

Lewis, W. and Eetemadi, S. (2013). Dramatically reducing training data size through vocabulary saturation. In *Proceedings of the Eigth Workshop on Statistical Machine Translation*, pages 281–291.

Lü, Y., Huang, J., and Liu, Q. (2007). Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350.

Mansour, S., Wuebker, J., and Ney, H. (2011). Combining translation and language model scoring for domain-specific data filtering. In *Proceedings of the International Workshop on Spoken Language TTranslation*, pages 222–229.

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 708–717, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mediani, M., Winebarger, J., and Waibel, A. (2014). Improving in-domain data selection for small in-domain sets. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT2014)*.

Moore, R. C. and Lewis, W. (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.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Peris, Á., Chinea-Rios, M., and Casacuberta, F. (2017). Neural networks classifier for data selection in statistical machine translation. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*.

Rousseau, A. (2013). Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82.

Wong, D. F., Lu, Y., and Chao, L. S. (2016). Bilingual recursive neural network based data selection for statistical machine translation. *Knowledge-Based Systems*, 108:15–24.