Analysing the Effect of Out-of-Domain Data on SMT Systems

Barry Haddow and Philipp Koehn
School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB, Scotland
{bhaddow,pkoehn}@inf.ed.ac.uk

Abstract

In statistical machine translation (SMT), it is known that performance declines when the training data is in a different domain from the test data. Nevertheless, it is frequently necessary to supplement scarce in-domain training data with out-of-domain data. In this paper, we first try to relate the effect of the out-of-domain data on translation performance to measures of corpus similarity, then we separately analyse the effect of adding the out-of-domain data at different parts of the training pipeline (alignment, phrase extraction, and phrase scoring). Through experiments in 2 domains and 8 language pairs it is shown that the out-of-domain data improves coverage and translation of rare words, but may degrade the translation quality for more common words.

1 Introduction

In statistical machine translation (SMT), domain adaptation can be thought of as the problem of training a system on data mainly drawn from one domain (e.g. parliamentary proceedings) and trying to maximise its performance on a different domain (e.g. news). There is likely to be some parallel data similar to the test data, but as such data is expensive to create, it tends to be scarce. The concept of "domain" is rarely given a precise definition, but it is normally understood that data from the same domain is in some sense similar (for example in the words and grammatical constructions used) and data from different domains shows less similarities. Data from the same domain as the test set is usually referred to as *in-domain* and data from a different domain is referred to as *out-of-domain*.

The aim of this paper is to shed some light on what domain actually is, and why it matters. The fact that a mismatch between training and test data domains reduces translation performance has been observed in previous studies, and will be confirmed here for multiple data sets and languages, but the question of why domain matters for performance has not been fully addressed in the literature. Experiments in this paper will be conducted on phrase-based machine translation (PBMT) systems, but similar conclusions are likely to apply to other types of SMT systems. Furthermore, we will mainly be concerned with the effect of domain on the translation model, since it depends on parallel data which is more likely to be in short supply than monolingual data, and domain adaptation for language modelling has been more thoroughly studied.

The effect of a shift of domain in the parallel data is complicated by the fact that training a translation model is a multi-stage process. First the parallel data is word-aligned, normally using the IBM models (Brown et al., 1994), then phrases are extracted using some heuristics (Och et al., 1999) and scored using a maximum likelihood estimate. Since the effect of domain may be felt at the alignment stage, the extraction stage, or the scoring stage, we have designed experiments to try to tease these apart. Experiments comparing the effect of domain at the alignment stage with the extraction and scoring stages have already been presented by (Duh et al., 2010), so we focus more on the differences between extraction and scoring. In other words, we examine whether adding more data (in or out-of domain) helps improve coverage of the phrase table, or helps improve the scoring of phrases.

A further question that we wish to address is

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 422–432,
Montréal, Canada, June 7-8, 2012.

whether adding out-of-domain parallel data affects words with different frequencies to different degrees. For example, a large out-of-domain data set may improve the translation of rare words by providing better coverage, but degrade translation of more common words by providing erroneous out-of-domain translations. In fact, the evidence presented in Section 3.5 will show a much clearer effect on low frequency words than on medium or high frequency words, but the total token count of these low frequency words is still small, so they don't necessarily have much effect on overall measures of translation quality.

In summary, the main contributions of this paper are:

- It presents experiments on 8 language pairs and 2 domains showing the effect on BLEU of adding out-of-domain data.
- It provides evidence that the difference between in and out-of domain translation performance is correlated with differences in word distribution and out-of-vocabulary rates.
- It develops a method for separating the effects of phrase extraction and scoring, showing that good coverage is nearly always more important than good scoring, and that out-of-domain data can adversely affect phrase scores.
- It shows that adding out-of-domain data clearly improves translation of rare words, but may have a small negative effect on more common words.

2 Related Work

The most closely related work to the current one is that of (Duh et al., 2010). In this paper they consider the domain adaptation problem for PBMT, and investigate whether the out-of-domain data helps more at the word alignment stage, or at the phrase extraction and scoring stages. Extensive experiments on 4 different data sets, and 10 different language pairs show mixed results, with the overall conclusion being that it is difficult to predict how best to include out-of-domain data in the PBMT training pipeline. Unlike in the current work, Duh et al. do not separate phrase extraction and scoring in order to analyse the effect of domain on them separately. They make the point that adding extra out-of-domain data may degrade translation by introducing unwanted lexical ambiguity, showing anecdotal evidence for this. Similar arguments were presented in (Sennrich, 2012).

A recent paper which does attempt to tease apart phrase extraction and scoring is (Bisazza et al., 2011). In this work, the authors try to improve a system trained on in-domain data by including extra entries (termed "fill-up") from out-of-domain data – this is similar to the `nc+epE` and `st+epE` systems in Section 3.4. It is shown by Bisazza et al. that this fill-up technique has a similar effect to using MERT to weight the in and out-of domain phrase tables. In the experiments in Section 3.4 we confirm that fill-up techniques mostly provide better results than using a concatenation of in and out-of domain data.

There has been quite a lot of work on finding ways of weighting in and out-of domain data for SMT (as opposed to simply concatenating the data sets), both for language and translation modelling. Interpolating language models using perplexity is fairly well-established (e.g. Koehn and Schroeder (2007)), but for phrase-tables it is unclear whether perplexity minimisation (Foster et al., 2010; Sennrich, 2012) or linear or log-linear interpolation (Foster and Kuhn, 2007; Civera and Juan, 2007; Koehn and Schroeder, 2007) is the best approach. Also, other authors (Foster et al., 2010; Niehues and Waibel, 2010; Shah et al., 2010) have tried to weight the input sentences or extracted phrases before the phrase tables are built. In this type of approach, the main problem is how to train the weights of the sentences or phrases, and each of the papers has followed a different approach.

Instead of weighting the out-of-domain data, some authors have investigated data selection methods for domain adaptation (Yasuda et al., 2008; Mansour et al., 2011; Schwenk et al., 2011; Axelrod et al., 2011). This is effectively the same as using a 1-0 weighting for input sentences, but has the advantage that it is usually easier to tune a threshold than it is to train weights for all input sentences or phrases. The other advantage of doing data selection is that it can potentially remove noisy (e.g. incorrectly aligned) data. However it will be seen later in this paper that out-of-domain data can usually contribute something useful to the translation system, so the 1-0 weighting of data-selection may be somewhat heavy-handed.

3 Experiments

3.1 Corpora and Baselines

The experiments in this paper used data from the WMT09 and WMT11 shared tasks (Callison-Burch et al., 2009; Callison-Burch et al., 2011), as well as OpenSubtitles data[1] released by the OPUS project (Tiedemann, 2009).

From the WMT data, both news-commentary-v6 (`nc`) and europarl-v6 (`ep`) were used for training translation models and language models, with `nc-devtest2007` used for tuning and `nc-test2007` for testing. The experiments were run on all language pairs used in the WMT shared tasks, i.e. English (en) into and out of Spanish (es), German (de), French (fr) and Czech (cs).

From the OpenSubtitles (`st`) data, we chose 8 language pairs – English to and from Spanish, French, Czech and Dutch (nl) – selected because they have at least 200k sentences of parallel data available. 2000 sentence tuning and test sets (`st-dev` and `st-devtest`) were selected from the parallel data by extracting every nth sentence, and a 200k sentence training corpus was selected from the remaining data.

Using test sets from both news-commentary and OpenSubtitles gives two domain adaptation tasks, where in both cases the out-of-domain data is europarl, a significantly larger training set than the in-domain data. The three data sets in use in this paper are summarised in Table 1.

The translation systems consisted of phrase tables and lexicalised reordering tables estimated using the standard Moses (Koehn et al., 2007) training pipeline, and 5-gram Kneser-Ney smoothed language models estimated using the SRILM toolkit (Stolcke, 2002), with KenLM (Heafield, 2011) used at runtime. Separate language models were built on the target side of the in-domain and out-of-domain training data, then linearly interpolated using SRILM to minimise perplexity on the tuning set (e.g. Koehn and Schroeder (2007)). Tuning of models used minimum error rate training (Och, 2003), repeated 3 times and averaged (Clark et al., 2011). Performance is evaluated using case-insensitive BLEU (Papineni et al., 2002), as implemented using the Moses `multi-bleu.pl` script.

[1] `www.opensubtitles.org`

Name	Language pairs	train	tune	test
Europarl (`ep`)	en↔fr	1.8M	n/a	n/a
	en↔es	1.8M	n/a	n/a
	en↔de	1.7M	n/a	n/a
	en↔cs	460k	n/a	n/a
	en↔nl	1.8M	n/a	n/a
News Commentary (`nc`)	en↔fr	114k	1000	2000
	en↔es	130k	1000	2000
	en↔de	135k	1000	2000
	en↔cs	122k	1000	2000
Subtitles (`st`)	en↔fr	200k	2000	2000
	en↔es	200k	2000	2000
	en↔nl	200k	2000	2000
	en↔cs	200k	2000	2000

Table 1: Summary of the data sets used, with approximate sentence counts

3.2 Comparing In-domain and Out-of-domain Data

The aim of this section is to provide both a qualitative and quantitative comparison of the three data sets used in this paper.

Firstly, consider the extracts from the English sections of the three training sets shown in Figure 1. The first extract, from the Europarl corpus, shows a formal style with long sentences. However this is still spoken text so contains a preponderance of first and second person forms. In terms of subject matter, the corpus covers a broad range of topics, but all from the angle of European legislation, institutions and policies. Where languages (e.g. English, French and Spanish) have new world and old world variants, Europarl sticks to the old world variants.

The extract from the News Commentary corpus again shows a formal tone, but because this is news analysis, it tends to favour the third person. It is written text, and covers a wider range of subjects than Europarl, and also encompasses both new and old world versions of the European languages.

The Subtitles text shown in the last example appears qualitatively more different from the other two. It is spoken text, like Europarl, but consists of short, informal sentences with many colloquialisms, as well as possible optical character recognition er-

Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural disasters that truly were dreadful.
You have requested a debate on this subject in the course of the next few days, during this part-session.
In the meantime, I should like to observe a minute' s silence, as a number of Members have requested, on behalf of all the victims concerned, particularly those of the terrible storms, in the various countries of the European Union.

(a) Europarl

Desperate to hold onto power, Pervez Musharraf has discarded Pakistan's constitutional framework and declared a state of emergency.
His goal?
To stifle the independent judiciary and free media.
Artfully, though shamelessly, he has tried to sell this action as an effort to bring about stability and help fight the war on terror more effectively.

(b) News commentary

I'il call in 30 minutes to check
Is your mother here, too?
Why are you outside?
It's no fun listening to women's talk
Well, why don't we go in together

(c) OpenSubtitles

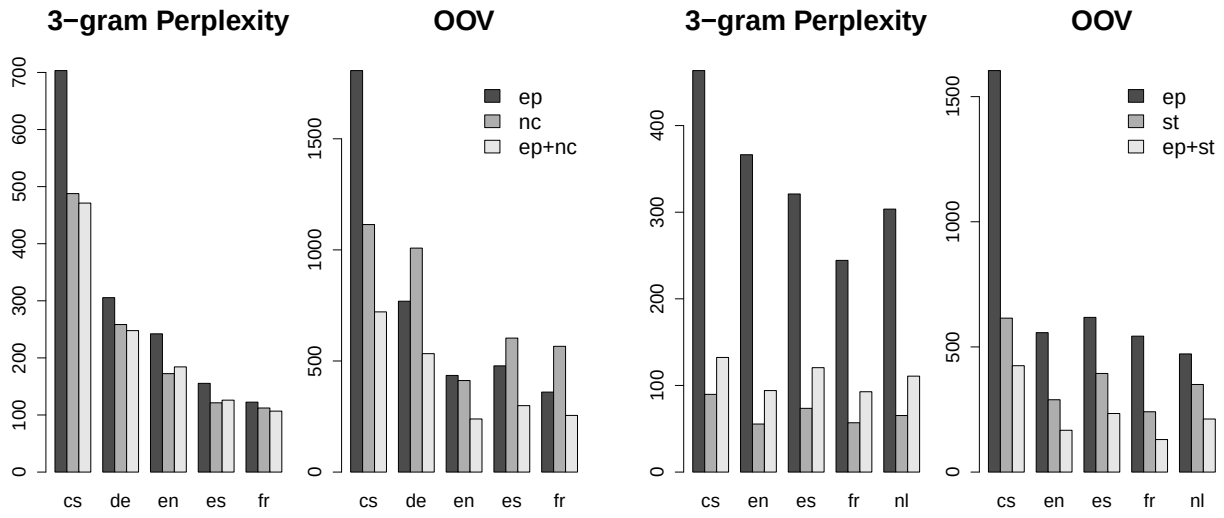
Figure 1: Extracts from the English portion of the training corpora

rors. It is likely to contain a mixture of regional variations of the languages, reflecting the diversity of the film sources.

In order to obtain a quantitative measure of domain differences, we used both language model (LM) perplexity, and out-of-vocabulary (OOV) rate, in the two test domains. For the `nc` domain, perplexity was compared by training trigram LMs (with SRILM and Kneser-Ney smoothing) on each of the `ep`, `nc` and `ep+nc` training sets, taking the intersection of the `ep` and `nc` vocabularies as the LM vocabulary. The perplexities of the `nc` test set wer calculated using each of the LMs. A corresponding set of LMs was trained to compare perplexities on the `st` test set, and all perplexity comparisons were performed on all five languages. The SRILM toolkit was also used to calculate OOV rates on the test set, by training language models with an open vocabulary, and using no unknown word probability estimation.

The perplexities and OOV rates on each test corpora are shown in Figure 2. The pattern of perplexities is quite distinct across the two test domains, with the perplexity from out-of-domain data relatively much higher for the `st` test set. The in-domain data LM also shows the lowest perplexity consistently on this test set, whilst for `nc`, the in-domain LM has a similar perplexity to the `ep+nc` LM. In fact for 3/5 languages (fr,cs and de) the `ep+nc` LM has the lowest perplexity.

With regard to the OOV rates, it is notable that for `nc` the rate is actually higher for the in-domain LM than the out-of-domain LM in three of the languages: French, German and Spanish. The most likely reason for this is that these languages have a relatively rich morphology, so the larger out-of-domain corpus (Table 1) gives greater coverage of the different grammatical suffixes. Czech shows a different pattern because in this case the out-of-domain corpus is not much bigger than the in-domain corpus, and English is morphologically much simpler so the increase in corpus size does not help the OOV rate so much.



(a) Test on news commentary.

(b) Test on subtitles.

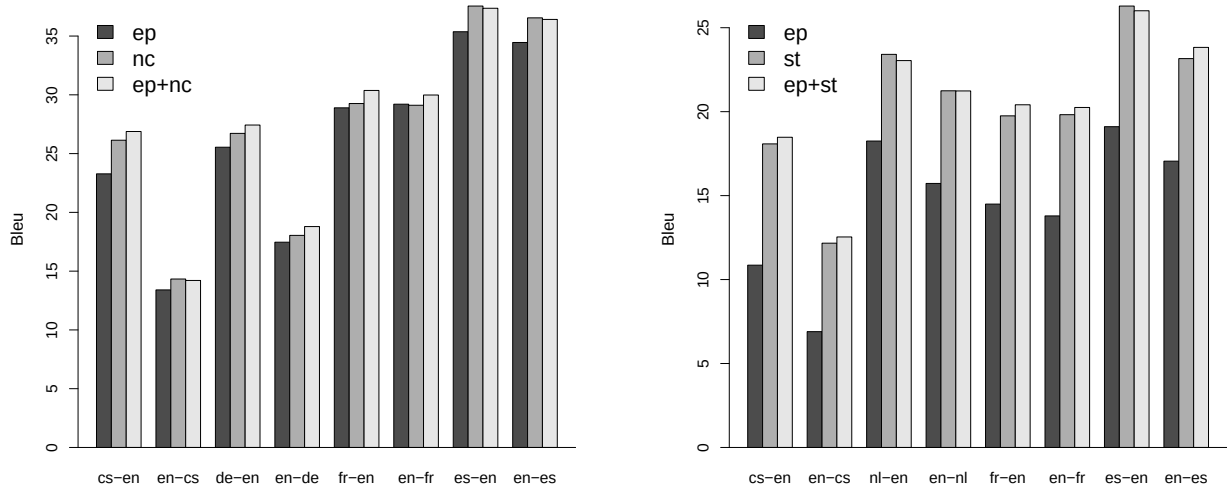
Figure 2: Comparison of perplexities and OOV rates on in-domain test data

3.3 Comparing Translation Performance of In and Out-of-domain Systems

Translation performance was measured on each of the test sets (`nc` and `st`) using systems built from just the in-domain parallel data, from just the out-of-domain parallel data, and on a concatenation of the in and out-of domain data. In other words, systems built from the `ep`, `nc` and `ep+nc` parallel texts were evaluated on the `nc` test data, and systems built from `ep`, `st` and `ep+st` were evaluated on the `st` test data. In all cases, the parallel training set was used to build both the phrase table and the lexicalised reordering models, the language model was the interpolated one described in Section 3.1, and the system was tuned on data from the same domain as the test set.

From Figure 3 it is clear that the difference between the in and out-of domain training sets is much bigger for `st` than for `nc`. The BLEU scores on `nc` for the `nc` trained systems are on average 1.3 BLEU points higher than those for the `ep` trained systems, whilst the scores on `st` gain an average of 6.0 BLEU points when the training data is switched from `ep` to `st`. The patterns are quite consistent across languages for the `st` tested systems, with the gains varying just from 5.2 to 7.2. However for the `nc` tested systems there are some language pairs which show a gain of more than 2 BLEU points when moving from out-of to in-domain training data (cs-en, en-es and es-en), whereas en-fr shows no change. The main link between the perplexity and OOV results in Figure 2 and the BLEU score variations in Figure 3 is that the larger in/out differences between the two domains is reflected in larger BLEU differences. However it is also notable that the two languages which display a rise in perplexity between `nc` and `ep+nc` are es and en, and for both es-en and en-es the `ep+nc` translation system performs worse than the `nc` trained system.

The BLEU gain from concatenating the in and out-of domain data, over just using the in-domain data can be quite small. For the `nc` domain this averages at 0.5 BLEU (with 3/8 language pairs showing a decrease), whilst for the `st` domain the average gain is only 0.2 BLEU (with again 3/8 language pairs showing a decrease). So even though adding the out-of-domain data increases the training set size by a factor of 10 in most cases, its effect on BLEU score is small.



(a) Test on news commentary

(b) Test on subtitles

Figure 3: Comparison of translation performance using models from in-domain, out-of-domain and joint data.

3.4 Why Does Adding Parallel Data Help?

In the previous section it was found that, across all language pairs and both data sets, adding in-domain data to an out-of-domain training set nearly always has a positive impact on performance, whilst adding out-of-domain data to an in-domain training set can sometimes have a small positive effect. In this section several experiments are performed with "intermediate" phrase tables (built from a single parallel corpus, augmented with some elements of the other parallel corpus) in order to determine how different aspects of the extra data affect performance. In particular, the experiments are designed to show the effect of the extra data on the alignments, the phrase scoring and the phrase coverage, whether adding in-domain data to an existing out-of-domain trained system, or vice-versa.

For each of the language pairs used in this paper, and each of the two domains, two series of experiments were run comparing systems built from a single parallel training set, intermediate systems, and systems built from a concatenation of the in and out-of-domain parallel data sets. Only the parallel data was varied, the language models were as described in Section 3.1, and the lexicalised reordering models were built from both training sets in all cases, except for the systems built from a single parallel data set[2]. This gives a total of four series of experiments, where the ordered pair of data sets (x,y) was set to one of (ep,nc), (nc,ep), (ep,st), (st,ep). In each of these series, the following translation systems were trained:

- x The translation table and lexicalised reordering model were estimated from the x corpus alone.
- x+y The translation system built from the x and y parallel corpora concatenated.
- x+yA As x but using the additional y corpus to create the alignments. This means that GIZA++ was run across the entire x+y corpus, but only the x section of it was used to extract and score phrases.
- x+yW As x+yA but using the phrase scores from the x+y phrase table. This is effectively the x+y system, with any entries in the phrase table that are just found in the y corpus removed.

[2]Further experiments were run using the parallel data from a single data set to build the translation model, and both data sets to build the lexicalised reordering model, but the difference in score compared to the x system was small (< 0.1 BLEU)

x+yE As x+yA but adding the extra entries from the x+y phrase table. This is effectively the x+y system, but with the scores on all phrases that are found in x phrase table set to their values from that table.

All systems were tuned and tested on the appropriate in-domain data set (either nc or st). Note that in the intermediate systems, the phrase table scores may no longer correspond to valid probability distributions, but this is not important as the probabilistic interpretation is never used in decoding anyway.

The graphs in Figure 4 show the performance comparison between the single corpus systems, the intermediate systems, and the concatenated corpus systems, averaged across all 8 language pairs. Table 2 shows the full results broken down by language pair, for completeness, but the patterns are reasonably consistent across language pair.

Firstly, compare the x+yW and x+yE systems, i.e. the systems where we add just the weights from the second parallel data set versus those where we add just the entries. When x is the out-of-domain (ep) data, then it is clearly more profitable to update the phrase-table entries than the weights from the in-domain data. In fact for the systems tested on st, the difference is quite striking with a $+5.7$ BLEU gain for the ep+stE system over the baseline ep system, but only a $+1.5$ gain for the ep+stW system. For the systems tested on the nc, adding the entries from nc gives a larger gain in BLEU than adding the weights ($+1.3$ versus $+0.8$), but both improve the BLEU scores over the ep+ncA system. The conclusion is that the extra entries from the in-domain data (the "fill-up" of Bisazza et al. (2011)) are more important than the improvements in phrase scoring that in-domain data may provide.

Looking at the other two sets of x+yW and x+yE systems, i.e. those where x is the in-domain data, tells another story. In this case, the results on both the nc and st test sets (Figure 4(b)) suggest that it is generally more useful to use the out-of-domain data as only a source of extra phrase-table entries. This is because the x+epE systems are the highest scoring in both cases, scoring higher than systems built from all the data concatenated by margins of 0.5 (for nc) and 0.4 (for st). This pattern is consistent across all the language pairs for nc, and across 5 of the 8 language pairs for st. Using the out-of-domain data set to update only the weights (the x+epW systems) generally degrades performance when compared to the systems that only use the ep data at alignment time (the x+epA systems).

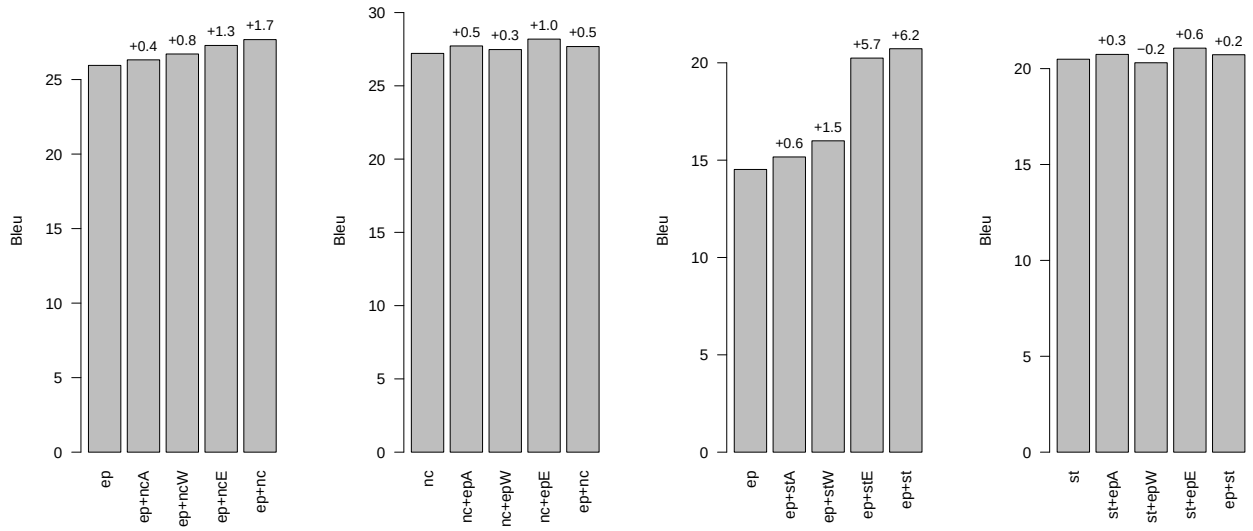
The size of the effect of adding extra data to the alignment stage only is mixed (as observed by (Duh et al., 2010)), but in general all the x+yA systems show an improvement over the x systems. In fact, for the st domain, adding ep at the alignment stage is the only consistent way to improve BLEU. Adding the weights, entries, or the complete out-of-domain data set does not always help.

3.5 Word Precision Versus Frequency

The final set of experiments addresses the question of whether the change of translation quality when adding out-of-domain has a different effect depending on word frequency. To do this, the systems trained on in-domain only are compared with the systems trained on all data concatenated, using a technique for measuring the precision of the translation for each word type.

To calculate the precision of a word type, it is necessary to examine each translated sentence to see which source words were translated correctly. This is done by recording the word alignment in the phrase mappings and tracking it through the translation process. If a word is produced multiple times in the translation, but occurs a fewer number of times in the reference, then it is assigned partial credit. Many-to-many word alignments are treated similarly. Precision for each word type is then calculated in the usual way, as the number of times that word appears correctly in the output, divided by the total number of appearances. The word types are then binned according to the log2 of their frequency in the in-domain corpus and the average precision for each bin calculated, then these are in turn averaged across language pairs.

The graphs in Figure 5 compare the in-domain source frequency versus precision relationship for systems built using just the in-domain data, and systems built using both in and out-of domain data. There is a consistent increase in precision for lower frequency words (occurring less than 30 times in training), but the total number of occurrences of these words is low, so they contribute less to over-

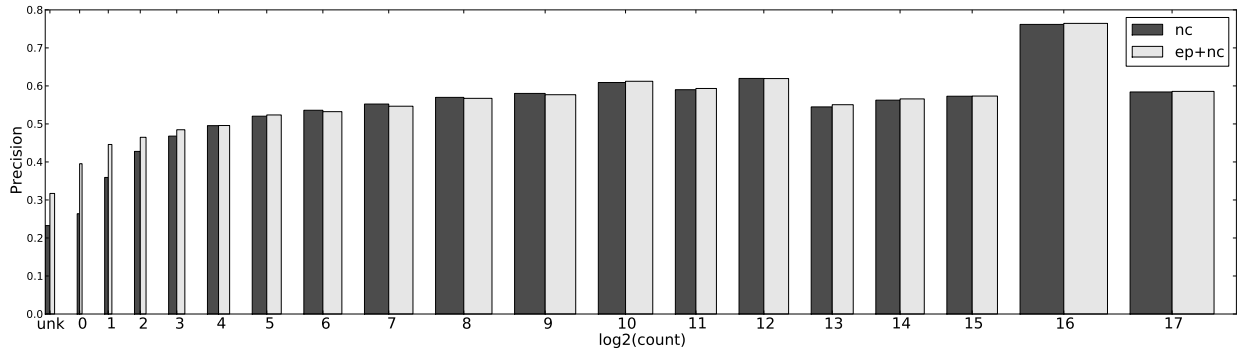


(a) Start with `ep`, test on `nc`. (b) Start with `nc`, test on `nc`. (c) Start with `ep`, test on `st`. (d) Start with `st`, test on `st`

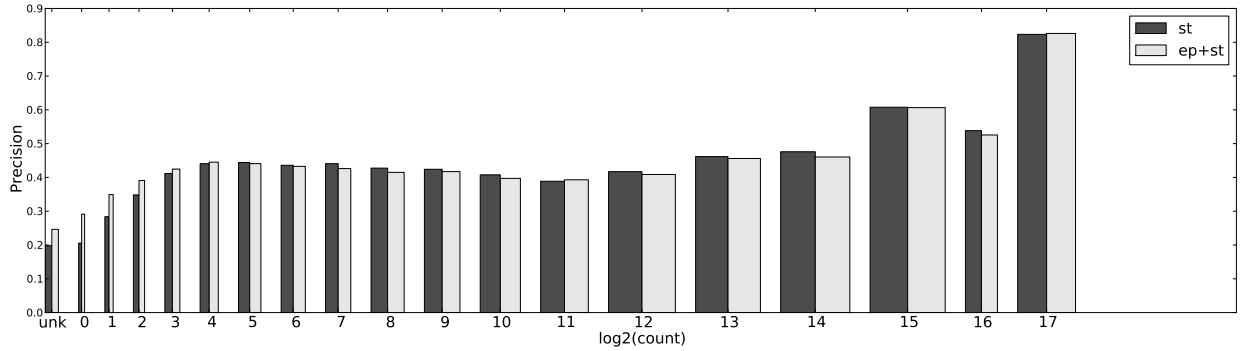
Figure 4: Showing the performance change when starting with either in or out-of domain data, and adding elements of the other data set. The "A" indicates that the second data set is only used for alignments, the "W" indicates that it contributes alignments and phrase scores, and the "E" indicates that it contributes alignments and phrase entries. The figures above each bar shows the performance change relative to the single corpus system.

System	cs-en	en-cs	de-en	en-de	fr-en	en-fr	es-en	en-es
ep	23.3	13.4	25.5	17.5	28.9	29.2	35.4	34.5
ep+ncA	23.5 (+0.2)	13.8 (+0.4)	25.9 (+0.4)	17.9 (+0.4)	29.3 (+0.4)	29.6 (+0.4)	35.7 (+0.3)	34.9 (+0.5)
ep+ncW	24.0 (+0.7)	14.2 (+0.8)	26.3 (+0.8)	18.2 (+0.7)	29.4 (+0.5)	29.8 (+0.6)	36.3 (+0.9)	35.6 (+1.1)
ep+ncE	26.2 (+2.9)	14.0 (+0.6)	27.0 (+1.5)	18.5 (+1.0)	29.7 (+0.9)	30.0 (+0.8)	37.0 (+1.7)	35.7 (+1.3)
nc	26.1 (+2.9)	14.3 (+0.9)	26.7 (+1.2)	18.0 (+0.6)	29.3 (+0.4)	29.1 (-0.1)	37.6 (+2.2)	36.5 (+2.1)
nc+epA	26.8 (+3.5)	14.6 (+1.2)	27.5 (+2.0)	18.5 (+1.0)	30.4 (+1.5)	29.9 (+0.7)	37.7 (+2.3)	36.4 (+2.0)
nc+epW	26.6 (+3.3)	14.4 (+1.0)	27.4 (+1.9)	18.4 (+1.0)	29.5 (+0.6)	29.8 (+0.6)	37.2 (+1.8)	36.5 (+2.0)
nc+epE	**27.4 (+4.1)**	**14.7 (+1.3)**	**28.1 (+2.6)**	**19.0 (+1.5)**	**30.9 (+2.0)**	**30.2 (+1.0)**	**38.4 (+3.0)**	**36.9 (+2.4)**
ep+nc	26.9 (+3.6)	14.2 (+0.8)	27.4 (+1.9)	18.8 (+1.3)	30.4 (+1.5)	30.0 (+0.8)	37.4 (+2.0)	36.4 (+2.0)
System	cs-en	en-cs	nl-en	en-nl	fr-en	en-fr	es-en	en-es
ep	10.9	6.9	18.2	15.7	14.5	13.8	19.1	17.1
ep+stA	11.9 (+1.0)	7.5 (+0.6)	19.0 (+0.8)	16.3 (+0.5)	15.0 (+0.5)	14.1 (+0.3)	19.8 (+0.7)	17.8 (+0.7)
ep+stW	12.2 (+1.3)	8.1 (+1.2)	20.0 (+1.7)	17.4 (+1.7)	15.8 (+1.3)	14.9 (+1.1)	20.8 (+1.7)	18.8 (+1.8)
ep+stE	18.0 (+7.1)	12.4 (+5.5)	22.5 (+4.2)	20.6 (+4.9)	19.6 (+5.1)	19.9 (+6.1)	25.6 (+6.5)	23.3 (+6.3)
st	18.0 (+7.2)	12.2 (+5.3)	23.4 (+5.1)	21.3 (+5.6)	19.7 (+5.2)	19.8 (+6.0)	26.3 (+7.2)	23.2 (+6.1)
st+epA	18.4 (+7.6)	12.4 (+5.5)	23.6 (+5.4)	21.3 (+5.6)	20.2 (+5.7)	20.1 (+6.3)	**26.4 (+7.3)**	23.5 (+6.5)
st+epW	18.2 (+7.3)	12.2 (+5.3)	22.4 (+4.2)	21.0 (+5.3)	19.9 (+5.4)	19.8 (+6.0)	25.8 (+6.7)	23.2 (+6.1)
st+epE	**19.1 (+8.3)**	12.5 (+5.6)	**24.0 (+5.8)**	**21.7 (+6.0)**	**20.6 (+6.1)**	**20.9 (+7.1)**	26.0 (+6.9)	23.7 (+6.6)
ep+st	18.5 (+7.6)	**12.5 (+5.6)**	23.0 (+4.8)	21.2 (+5.5)	20.4 (+5.9)	20.2 (+6.5)	26.0 (+6.9)	**23.8 (+6.8)**

Table 2: Complete scores for the experiments described in Section 3.4 and summarised in Figure 4. Naming of the systems is explained in the text, and in the caption for Figure 4



(a) News commentary



(b) Subtitles

Figure 5: Performance comparison of in-domain systems versus systems built from in and out-of domain data concatenated. Precision is plotted against log2 of in-domain training frequency, and averaged across all 8 language pairs. The width of the bars indicates the average total number of occurrences in the test set.

all measures of translation quality. For the words with moderate training set frequencies, the precision is actually slightly higher for the systems built with just in-domain data, an effect that is more marked for the `st` domain.

4 Conclusions

In this paper we have attempted to give an in-depth analysis of the domain adaptation problem for two different domain adaptation problems in phrase-based MT. The differences between the two problems are clearly illustrated by the results in Figures 2 and 3, where we see that the difference between the in-domain and out-of-domain data are larger for the OpenSubtitles domain than for the News-Commentary domain. This can be detected by the differences in word distribution and out-of-vocabulary rates observed in Figure 2, and is reflected by the differing translation results in Figure 3.

However, the experiments of Sections 3.4 and 3.5 show some common themes emerging in the two domains. In both cases, the out-of-domain data helps most when it is just allowed to add entries (i.e. "fill in") the phrase-table, and using the scores provided by out-of-domain data has a tendency to be harmful to translation quality. The precision results of Section 3.5 show out-of-domain data (when it is simply added to the training set) mainly helping with the low frequency words, and having a neutral or harmful effect for higher frequency words. This explains why approaches which try to weight the out-of-domain data in some way (e.g. corpus weighting or instance weighting) can be more successful than

simply concatenating data sets. It also suggests that the way forward is to look for methods that use the out-of-domain data mainly for rarer words, and not to change translations which have a lot of evidence in the in-domain data.

5 Acknowledgments

This work was supported by the EuroMatrixPlus[3] and Accept[4] projects funded by the European Commission (7th Framework Programme).

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *Proceedings of IWSLT*.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1994. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Jorge Civera and Alfons Juan. 2007. Domain Adaptation in Statistical Machine Translation with Mixture Modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of ACL*.

[3] www.euromatrixplus.net

[4] www.accept.unige.ch

Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Analysis of translation model adaptation in statistical machine translation. In *Proceedings of IWSLT*.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June. Association for Computational Linguistics.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA, October. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Philipp Koehn and Josh Schroeder. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL Demo Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Saab Mansour, Joern Wuebker, and Hermann Ney. 2011. Combining Translation and Language Model Scoring for Domain-Specific Data Filtering. In *Proceedings of IWSLT*.

Jan Niehues and Alex Waibel. 2010. Domain Adaptation in Statistical Machine Translation using Factored Translation Models. In *Proceedings of EAMT*.

Franz J. Och, Christoph Tillman, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Franz J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th*

Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Holger Schwenk, Patrik Lambert, Loïc Barrault, Christophe Servan, Sadaf Abdul-Rauf, Haithem Afli, and Kashif Shah. 2011. LIUM's SMT Machine Translation Systems for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 464–469, Edinburgh, Scotland, July. Association for Computational Linguistics.

Rico Sennrich. 2012. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of EACL*.

Kashif Shah, Loïc Barrault, and Holger Schwenk. 2010. Translation Model Adaptation by Resampling. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 392–399, Uppsala, Sweden, July. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing, vol. 2*, pages 901–904.

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (vol V)*, pages 237–248. John Benjamins, Amsterdam/Philadelphia.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of Selecting Training Data to Build a Compact and Efficient Translation Model. In *Proceedings of IJCNLP*.