Analysis of Translation Model Adaptation in Statistical Machine Translation

Kevin Duh, Katsuhito Sudoh, Hajime Tsukada

NTT Communication Science Laboratories
2-4 Hikari-dai, Seika-cho, Kyoto 619-0237, JAPAN
{kevinduh,sudoh,tsukada}@cslab.kecl.ntt.co.jp

Abstract

Numerous empirical results have shown that combining data from multiple domains often improve statistical machine translation (SMT) performance. For example, if we desire to build SMT for the medical domain, it may be beneficial to augment the training data with bitext from another domain, such as parliamentary proceedings. Despite the positive results, it is not clear exactly *how* and *where* additional out-of-domain data helps in the SMT training pipeline. In this work, we analyze this problem in detail, considering the following hypotheses: out-of-domain data helps by either (a) improving word alignment or (b) improving phrase coverage. Using a multitude of datasets (IWSLT-TED, EMEA, Europarl, OpenSubtitles, KDE), we show that sometimes out-of-domain data may help word alignment more than it helps phrase coverage, and more flexible combination of data along different parts of the training pipeline may lead to better results.

1. Introduction

The performance of Statistical Machine Translation (SMT) critically depends on the *quality and quantity* of the input training data. However, training data in the form of bitext (i.e. sentence-aligned corpora) can be expensive to obtain. One solution is domain adaptation. For example, suppose we are interested in building SMT for the medical domain, but have only little bitext for medical-related documents (i.e. insufficient *in-domain* data). Domain adaptation methods seek to improve performance by adding data from other domains/genres, such as parliamentary proceedings (i.e. *out-of-domain* data).

Research in domain adaptation can be considered as finding an optimal tradeoff between quality and quantity. The overall quantity of data is increased by adding out-of-domain data; but from the perspective of translating in-domain (medical) text, out-of-domain (parliamentary) data may be of lower quality. Therefore, while numerous empirical results have shown that combination of in-domain and out-of-domain data often achieve improvements, we believe it is important to analyze the results in more detail. Here, we seek to discover exactly *how* and *where* out-of-domain data helps in the SMT training pipeline.

Our analysis focuses on the translation model training pipeline under the phrase-based SMT framework [1, 2]. The pipeline can be viewed generally as a two-step procedure:

1. Word alignment: given sentence-aligned bitext, find the individual word correspondences.

2. Phrase extraction/scoring: given word alignment results, find the phrase pairs and their translation probabilities.

Out-of-domain data can benefit either, both, or neither of these steps. We perform experiments on multiple domain adaptation problems (involving the IWSLT, EMEA, Europarl, KDE, and OpenSubtitles datasets) to gauge the influence of out-of-domain data in each of these two steps. Our findings can be summarized as follows:

- Out-of-domain data can be a "double-edged sword": sometimes adding it helps; other times it hurts. Importantly, its effect may differ substantially at different steps in the pipeline.

- There are scenarios in which out-of-domain data mainly benefits the word alignment step. In this case better results may be obtained by including out-of-domain data for word alignment training, but *excluding* it during phrase extraction. The reason this happens is that out-of-domain data can help decrease the lexical translation ambiguity of in-domain words, but otherwise does *not* improve out-of-vocabulary rate.

- In general, our experiments on 10 language pairs show that it is difficult to predict where out-of-domain data will be helpful *a priori*. Thus, as a practical guideline, we advocate performing model selection on the various training methods presented in this paper.

The paper is organized as follows: First, we explain in Section 2 our analysis methodology, i.e. how we design the experiments in order to analyze the effects of out-of-domain data. Then, Section 3 presents in-depth analysis on one dataset, the IWSLT 2010 TED Talk Translation task. Thereafter, extensive experiments involving ten language pairs for EMEA, Europarl, KDE, and OpenSubtitles are presented in Section 4. Finally, we end with discussions on related work (Section 5) and conclusions (Section 6).

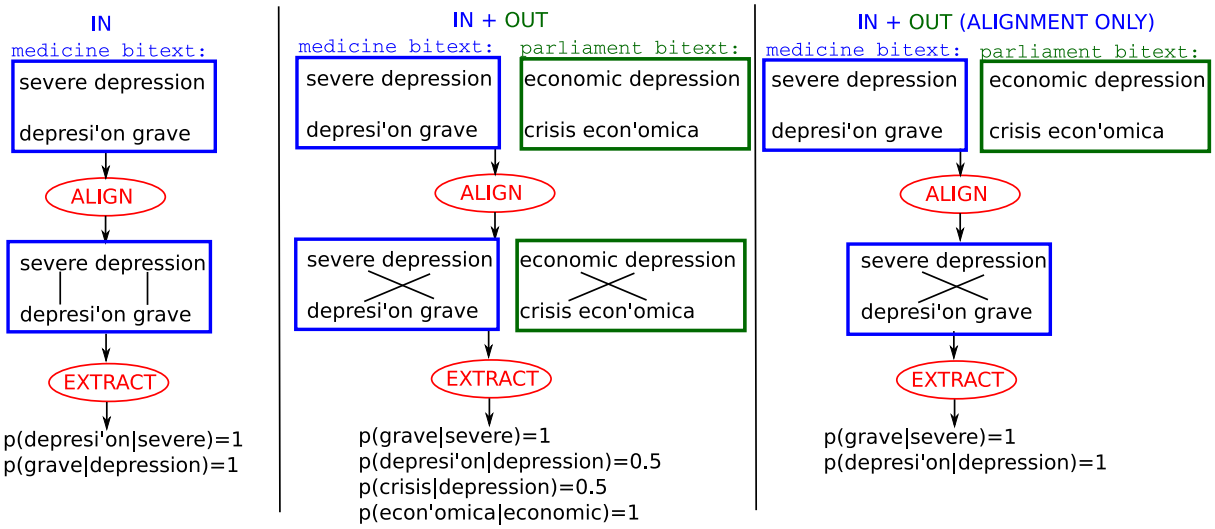

Figure 1: Experiment Design: To measure the effect of out-of-domain (parliament) data on word alignment and phrase coverage, we compare SMT performance of the following systems. System **IN** trains on only in-domain (medicine) data; System **IN+OUT** trains on concatenated in-domain and out-of-domain data. If **IN+OUT** outperforms **IN**, then out-of-domain data helped, but we cannot attribute the cause to better alignment or phrase coverage. System **IN+OUT (Align Only)** uses concatenated data for word alignment but only in-domain text for phrase extraction: thus OOV rate stays constant and improvements can be attributed to better alignments.

2. Analysis Methodology

2.1. Main Idea

The goal of this work is to understand where out-of-domain data might help in the training pipeline. We have the following working hypotheses:

1. *Out-of-domain data improves word alignment:* For instance, the alignment of infrequent words are often difficult to predict correctly; additional out-of-domain data can help reduce the lexical translation ambiguity.

2. *Out-of-domain data improves phrase coverage:* Words or phrases that are not seen in the training data cannot be translated. Additional out-of-domain data, by its sheer quantity, may reduce the number of such out-of-vocabulary (OOV) words or phrases. Note that coverage can improve for either the input (e.g. Foreign) or output (e.g. English); here we refer to both and do not make a distinction.

Our goal is to tease apart the effect of out-of-domain data on word alignment and phrase coverage when adding it helps overall SMT performance (e.g. BLEU score). To do this, we design the following experiment setup: First, we divided the training pipeline into two steps, word alignment and phrase extraction/scoring. Then, we run separate experiments where out-of-domain data is inserted at different parts of the pipeline. Finally, we infer how out-of-domain data helped by measuring the respective BLEU scores of the final models.

Figure 1 presents our experimental design in detail. For exposition purposes, throughout this section, we assume medicine bitext is in-domain, and parliamentary proceedings is out-of-domain. We compare several systems:

- **IN**: Only the in-domain data is used, and it is run through the conventional training pipeline. This is the no-adaptation case.

- **IN+OUT**: In-domain and out-of-domain data are concatenated, then sent through the conventional training pipeline. This is a straightforward and often effective method for domain adaptation. However, if we observe improvements, it is difficult to attribute the cause to better word alignment or better phrase coverage.

- **IN+OUT (Alignment Only)**: This is a novel setup where in-domain and out-of-domain data are concatenated for word alignment, but phrase extraction is only performed on the in-domain portion. Since the out-of-domain data is not involved in phrase extraction (and thus OOV rate does not change), we can attribute the improvements to better alignments.

In addition to the above systems, we also consider two more systems, **IN+IR** and **IN+IR (Alignment Only)**: these systems add a focused subset of the out-of-domain data, which is obtained by information retrieval (IR) methods [3, 4]. Adding the entire out-of-domain data might be risky; IR methods reduce this risk by adding only those bitext that are "similar" to in-domain data. It is another effective

method for domain adaptation, and provides an interesting comparison to **IN+OUT** and **IN+OUT (Alignment Only)**.[1]

Our IR method is implemented as follows: first, we collect all ngrams in the *in-domain training data* for both sides of the language pair. For each language side, a hash is created where the key represents the ngram and the value represents the count of this ngram in the training data. Then, an out-of-domain sentence is selected if it contains an ngram in this hash, and the hash value is decremented. We no longer retrieve matches if the hash value becomes zero, similar to the Joshua subsampling technique [5]. The rationale for this is to have a balanced coverage of ngrams. This procedure is performed independently for each language side, and the union of the selected sentences forms the IR bitext (i.e. we take the sentence pair if at least one side is retrieved).[2]

2.2. Intuition and Example Sentences

To establish some intuition before we present the experiments, let us look as some real data. The following are two English-Spanish bitexts from medical text (in-domain) and parliament proceedings (out-of-domain):

Medicine (EMEA):

`if you have severe depression, you must not use avonex . / no debe utilizar avonex si padece una depresión grave .`

Parliament (Europarl):

`the economic depression in europe has lasted at least ten years . / europa sufre una crisis económica desde hace , al menos , diez años .`

Where might adding parliamentary data help, and where might it hurt? Focusing on the underlined words, note that the correct translation for the English word "depression" depends on the domain. In medicine, it refers to mental depression so is translated as "depresión"; but when it refers to economic depression (as is common in Europarl), it is often translated as "crisis". This is an example where simply adding out-of-domain might *degrade* performance. For **IN+OUT**, we simply increase the risk of translating "depression" to "crisis", which is inappropriate for medicine text.

On the other hand, it is entirely possible that word alignment is *improved*: if the phrase "severe depression" is infrequent in the medicine data, it may be difficult to learn whether "severe" aligns to "grave" or "depresión" (i.e. both may have equal likelihood in EM training). The Europarl data, fortunately, contains many examples where "severe" and "grave" co-occur, which decreases the lexical ambiguity. Therefore, the **IN+OUT (Alignment Only)** system may get better word alignments, without the risk of introducing spurious out-of-domain words in the translation model.

On the other hand, here is an example of out-of-domain data improving coverage:

Parliament (Europarl):

`the hague conference on climate change / conferencia de la haya sobre el cambio climático`

Proper names like "the hague" have the same translation "la haya" regardless of domain. It is an OOV item for the medicine training data, so adding this parliamentary sentence clearly improves phrase coverage.

2.3. Clarifications

One point needs to be clarified. A key distinction made in this work is that improvements can come from better word alignment or better phrase coverage. However, even though we say that "word alignment improves", what really impacts the final SMT performance is still the phrases that are extracted from these word alignments. Therefore, we want to emphasize that the improvements really come from (1) *better in-domain phrases*, due to better word alignments, and (2) *better out-of-domain phrases*, due to better coverage.

Also, we note that "better word alignments" does not necessary mean better F-score or alignment error rate (AER), since the relation between these metrics and BLEU is not linear [6, 7]. Here we use an operational definition and say that word alignment improved if the final BLEU score of **In+Out (Alignment Only)** outperforms **In**, since word alignment is the only step that differs in the two systems.

3. Experiments on IWSLT TED Talks

3.1. Corpora and Setup

As an in-depth evaluation, we use data provided by the IWSLT 2010 TED Talks Challenge Evaluation. The task is English-to-French translation of speeches on a variety of topics (mostly on global issues, technology, culture, and design). This in-domain corpus consist of 300+ talks; we divided it into training, dev, and test set, with 84k, 644, and 663 sentences respectively. The out-of-domain data comes from WMT2010 and consists of Europarl (1.2M sentences), News Commentary (67k sentences), and United Nations (4.9M sentences) parallel corpus. These differ from in-domain in both content and style. Our IR approach (using 4grams) extracted 142k sentences from out-of-domain corpora.

Using standard language model adaptation, we used SRILM to interpolate 4gram models trained on the French side of all available bitext (optimized on dev perplexity). For each system, we rerun minimum error rate training on the dev set to re-tune optimal weights. Translation model building and decoding uses the Moses software, with a full lexical reordering and grow-diag-final-and phrase extraction heuristic. Text is lower-cased and punctuation is kept.

[1] For notational simplicity, we sometimes refer to **IN+OUT (Alignment Only)** as in+out(A) or i+o(A) when the meaning is clear in context. Similarly, in+ir(A) means **IN+IR (Alignment Only)**.

[2] Note that this method differs from previous IR approaches [3, 4] in two important aspects: (1) we sample using the training data, not the test data; (2) we sample on both sides of the language pair, rather than just the input side. We think working with the training data is a more practical solution, and also opens up the possibility to perform novel IR techniques using both languages jointly.

3.2. Main Results

Table 1 presents the BLEU results on the test data. In general, adding out-of-domain data improves results. The best result is achieved by **IN+IR(A)** (23.28), a 1.24 improvement over the baseline **IN** (22.04). Importantly, **IN+IR(A)** also outperformed **IN+IR** (22.66) by 0.6 points–thus, *excluding* the out-of-domain data for phrase extraction actually led to more improvements. This supports our claim that out-of-domain data has different effects on different parts of the training pipeline.

It is also interesting to compare the results between **IN+IR** (22.66) and **IN+OUT** (22.83) . Adding all of the out-of-domain data appears to benefit slightly more, possibly due to a decrease of OOV rate from 1.5% to 0.8%. On the other hand **IN+OUT(A)** underperforms **IN+OUT**, which is a different trend compared to **IN+IR(A)** outperforming **IN+IR**. Thus, it is not always the case that increasing out-of-domain data will necessarily improve word alignment.

To understand the results, we compute various statistics of the trained systems, especially comparing **IN** with the alignment-only systems (**IN+IR(A)**, **IN+OUT(A)**). We observe that alignment-only systems have *fewer* alignment points, leading to a larger phrase table[3] and smaller lexical translation tables. e.g. for **IN+OUT(A)**, 52% of lexical entries had smaller ambiguity (fewer translation options) compared to that of **IN**. This means that adding out-of-domain data for alignment training, while restricting alignment inference/decoding to in-domain data, can reduce the number of translation options for a given word. We think this is the major effect at play when alignment-only systems do well.

3.3. Detailed Analysis of BLEU Improvements

The BLEU scores indicate that **IN+IR(A)** is better than **IN** as well as **IN+IR**. We now dissect the BLEU scores to find exactly where the improvements come from.

As an analysis technique, we perform a two-way comparison between ngrams generated by two systems, with respect to the reference. First, for each reference sentence, we list all the ngrams from the **IN** system that have a match; similarly we list all the correct ngrams from **IN+IR(A)** output. In the BLEU calculation, the percentage of the correct ngrams is used to compute precision. Here, instead we look at the identities of the ngrams and ask: which ones are shared between **IN** and **IN+IR(A)**, and which ones are unique to one system? By looking at the ngrams that are unique, we can trace the source of the improvement.

First, we compare **IN** and **IN+IR(A)**: **IN** has 6298 correct ngrams out of 21849, and **IN+IR(A)** has 6489 correct ngrams out of 21681. Among these, 637 correct ngrams are unique to **IN+IR(A)**. We are interested to see how many of these unique ngrams are *not present in the phrase table* of **IN**; this represents the improvement due to new in-domain phrases extracted from better alignments. The result: 40% of the unique correct ngrams in **IN+IR(A)** is due to alignment alone, while the remaining 60% could have potentially be generated by the original **IN** system.

Second, we compare **IN+IR(A)** and **IN+IR**. **IN+IR** has 6355 out of 21529 ngrams correct, and there are 610 correct ngrams that is unique to **IN+IR**. We are interested to see how many of ngrams are novel with respect to the **IN+IR(A)** phrase table; these represent new out-of-domain phrases that helped reduced OOV rate. The result: only 28% of these ngrams do not occur in the **IN+IR(A)** phrase table. Thus, the effect of improved OOV rate appears to be rather minimal.

Finally, we look at the *incorrect* ngrams that are generated. Both **IN+IR(A)** and **IN+IR** systems generate roughly 15k incorrect ngrams. Among these incorrect ngrams, 5874 are unique to **IN+IR**. We are interested to see how many of these are novel with respect to the **IN+IR** bitext (not phrase table). This represents incorrect translations due to spurious out-of-domain words that irrelevant for the in-domain task. The result: 68% of these unique ngrams are not seen anywhere in the in-domain data, so it appears that a large fraction of mistakes can be attributed to the introduction of inappropriate out-of-domain phrases.

To summarize, by looking at the identities of the ngrams that are unique to one system, we conclude that:

- An system that uses out-of-domain data to improve alignment can generate considerable number of new in-domain phrases, which lead to BLEU improvements. (40% of cases).
- Many good ngrams generated by **IN+IR** are also contained by the phrase table of an in-domain system (72% of cases). Thus the effect of reduced OOV rate is relatively small (this may be because the original OOV rate is already low in this task).
- Many of the mistakes made by **IN+IR** is due to extraneous translation options introduced by out-of-domain text (68% of cases). These mistakes would not occur if phrase extraction were only performed in-domain.

3.4. Example Alignment

We manually inspected the sentences to see how out-of-domain data changes word alignments. It appears that alignment improvements often occur at rare words.

For example, consider the alignment matrix in Table 2, focusing on the phrase pair “world war ii”. System **IN+IR(A)** achieves the correct alignment (“world”∼”mondiale”, ”war”∼”guerre”, ”ii”∼”seconde”), but **IN** does not (e.g. “ii”∼”mondiale”). It turns out this is due to the low count of “ii” in the TED corpus, occurring only 22 times, almost all in the context of “world war ii”. However, in French, this is not always translated as “seconde guerre mondiale”, but sometimes as “2éme guerre mondiale”. This makes the co-occurrence of “ii” and “mondiale” higher than that of “ii” and “seconde”. For a larger (out-of-domain) training data, the problem disappears because

[3]Fewer alignment points lead to large phrase tables because the extraction heuristic is able to find more consistent phrase pairs [8].

	in	in+out	i+o(A)	in+ir	in+ir(A)
Test BLEU	22.04	22.83	22.43	22.66	**23.28**
Training Set Size for Alignment (# sentence pairs)	83.9k	6409k	6409k	307k	307k
Training Set Size for Extraction (# sentence pairs)	83.9k	6409k	83.9k	307k	83.9k
Avg # Alignment Points per sentence	11.46	21.51	11.08	21.61	11.19
Phrasetable size (# entries)	1.88M	202M	1.97M	15.7M	1.94M
Lexical translation table size (# entries, lex.e2f)	213k	4885k	170k	802k	185k
% lexical entries w/ larger ambiguity than (in)	-	83%	10%	67%	13%
% lexical entries w/ smaller ambiguity than (in)	-	5%	52%	8%	42%
Out-of-vocabulary rate (%)	2.50	0.80	2.10	1.50	2.30

Table 1: IWSLT Results and Statistics for five systems. See Section 2.1 for system descriptions. in+ir(A) achieves the best BLEU; the fact that it outperforms in+ir especially indicates that word alignment is more important than phrase coverage for this task. The statistics indicate that adding out-of-domain data only for alignment leads to (1) decrease in number of alignment points, (2) increase in phrase table size, (3) decrease in lexical translation table size and translation ambiguity.

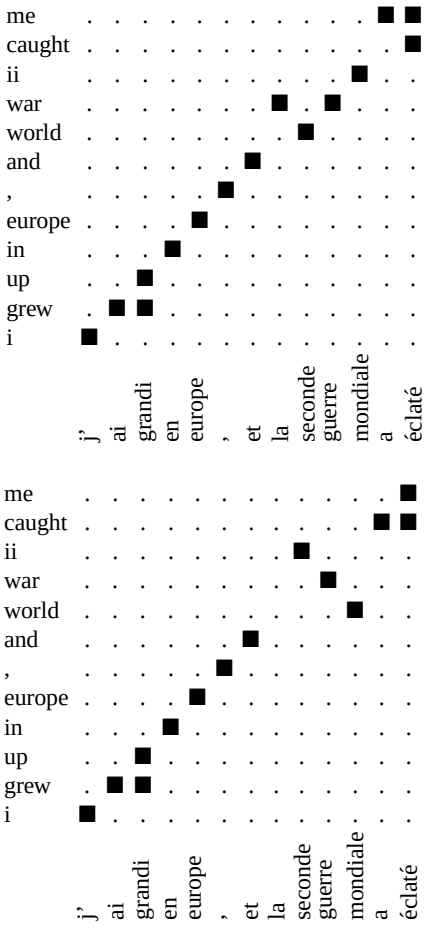

Table 2: Word Alignments of [top] System **IN** vs. [bottom] **IN+IR(A)**.

words like "mondiale" (whose translation rarely changes by domain) has more reliable co-occurrence statistics.

In general, we find that adding out-of-domain data for alignment training, but excluding it for phrase extraction, has the tendency to sharpen the lexical translation probabilities. This may be a good thing for small in-domain datasets because the original lexical translation probabilities are usually too ambiguous due to low co-occurrence counts.

4. Experiments on Ten Language Pairs

The IWSLT experiments in Section 3 demonstrated that using out-of-domain data only for word alignment is better than using it for all steps in the training pipeline. We now present results for more tasks and language pairs. To give a sneak preview: the results over 4 tasks and 10 language pairs (resulting in 240 SMT systems) show that the best condition is more varied: *no single system emerges as the winner in all cases.* Although this is not a satisfying conclusion, it is in a sense not surprising that increasing the number of experiments also increases the variance of the result. We attempt to perform some analysis and discover some guidelines for domain adaptation in a wide variety of scenarios.

4.1. Corpora and Setup

We experiment with corpora in four different domains: Europarl (parliamentary proceedings) is from [9]; EMEA (medical text), OpenSubtitles (movie subtitles), and KDE (computer localization files) are from the OPUS collection [10]. Europarl data is the largest; OpenSubtitles and KDE are the smallest. The statistics of the training data are shown in Table 3. For most cases, we reserved 2000 sentences for development and test, respectively; for datasets smaller than 70k sentences, we hold-out 700 sentences instead.

We setup four domain adaptation tasks, shown in Table 4. Each task consists of ten language pairs (focusing on translation into English). For all systems, the language model is a

SRILM interpolated 3gram LM. Minimum error rate training on the dev set is run independently for each system. Translation model training is done with Moses, using full lexical reordering and grow-diag-final-and heuristic.

Task		Statistics		
in	out	#IR sent	OOV(in)	OOV(+out)
EMEA	Europarl	197k	2.87	2.39
Europarl	EMEA	127k	0.62	0.57
OpenSub	Europarl	109k	7.44	3.41
KDE	Europarl	37k	6.61	4.34

Table 4: Four domain adaptation tasks, with some statistics: (1) #IR sent=number of out-of-domain sentences extracted by IR method. (2) OOV(in) and OOV(+out) are the OOV rates for the in-domain and combined in+out data.

4.2. Main Results

The BLEU results for all tasks are shown in Tables 5, 6, 7, and 8. Each row in the table compares 6 systems for a language pair (e.g. da→en). The best system in each row, as well as any other system within 0.20 BLEU score, are boldfaced. We consider these boldfaced numbers as the best system(s) for an experiment. Note that different systems win under different tasks or language pairs.

Since there are many numbers in the results tables, we present some summary statistics for each task: (1) the row $\star$ counts the number of times a system performs best; (2) the row $\Diamond$ indicates whether the difference between a pair of systems is "significant".[4]

Our observations for each task are:

- **OpenSub: IN+IR(A)** frequently outperforms **IN** (7 of 10 times), suggesting out-of-domain data provides better alignments. Other results are less consistent.
- **Europarl:** There is a trend showing that **IN+IR(A)** and **IN+IR** are best (though differences are not large). This shows that even small amounts of out-of-domain data may still help.
- **EMEA:** Adding out-of-domain Europarl often hurts results. We found that the EMEA corpus contains many repeated phrases (e.g. "keep out of the reach of children", a common disclaimer in many medical products). This may make it an easier task, requiring less training data.
- **KDE:** Out-of-domain data certainly helps, but there is no clear winner among **IN+IR**, **IN+IR(A)**, and **IN+OUT(A)**.

[4]The first statistic counts the best out of 6 system, whereas the second statistic looks at all pairwise comparisons between systems. The latter is determined by counting how many times system A outperforms system B; if A outperforms B more than 7 times in 10 language pairs, we indicate it in row $\Diamond$ as A $>$ B. Naturally, this is not a strict statistical significance test, but we think it gives some general insights.

	in	out	in+out	i+o(A)	in+ir	in+ir(A)
da	**29.29**	28.22	24.36	21.38	21.06	21.54
de	17.22	14.90	**18.04**	17.40	16.34	17.37
el	25.20	23.92	28.08	**29.13**	28.07	27.05
es	35.60	22.14	34.27	**35.82**	35.00	**35.66**
fi	9.91	8.06	**10.92**	10.31	10.59	10.23
fr	**11.77**	7.78	11.25	11.30	11.05	11.23
it	20.40	19.89	21.60	20.17	**22.12**	20.69
nl	12.79	8.30	12.50	**12.95**	**13.03**	**12.98**
pt	30.22	22.22	30.21	24.80	30.54	**30.89**
sv	**36.87**	26.35	35.48	36.39	36.48	**36.77**
$\star$	3	0	2	3	2	4
$\Diamond$	in+ir(A)>in					

Table 5: OpenSubtitles task: test BLEU.

	in	out	in+out	i+o(A)	in+ir	in+ir(A)
da	**29.54**	14.45	29.36	29.29	**29.50**	**29.58**
de	**27.03**	12.20	26.84	26.82	**27.01**	**27.00**
el	27.70	11.20	27.71	27.72	27.50	**27.93**
es	30.89	10.48	30.77	**31.03**	**31.17**	**31.02**
fi	**25.20**	6.96	24.84	25.06	**25.24**	**25.29**
fr	30.11	14.12	**30.31**	30.02	**30.21**	**30.32**
it	**27.30**	12.86	**27.35**	**27.21**	**27.22**	**27.24**
nl	23.04	12.87	24.95	**25.19**	**25.18**	**24.99**
pt	31.04	15.22	**31.17**	**31.27**	31.06	**31.12**
sv	**32.93**	16.56	**33.03**	**33.22**	32.98	**33.10**
$\star$	5	0	4	5	7	10
$\Diamond$	in+ir(A)>{in,in+out,in+ir}					

Table 6: Europarl task: test BLEU

4.3. Predicting when a method performs well

Given the variety of results, we attempted some meta-analysis to see if we can predict when a particular method will outperform another, using features about the dataset. In particular, we consider what features result in **IN+IR(A)** outperforming **IN+IR**, by framing a binary classification problem between the two. Each row in Tables 5, 6, 7, and 8 represents a sample. We extract the six features per sample:

1. Number of in-domain words
2. OOV rate of **IN**
3. OOV improvement by using **IN+IR**
4. Number of retrieved words
5. Fraction of retrieved words over in-domain words
6. Baseline BLEU score of **IN**

A linear SVM classifier was trained on the 40 samples (4 tasks, 10 language pairs) in order to discover the features that are predictive of performance. The 10-fold cross-validation accuracy is 67.5%. We observe that the features that are positively correlated with **IN+IR(A)** outperforming **IN+IR** are

	EMEA		Europarl		OpenSubtitles		KDE	
	#sent	#word(for/en)	#sent	#word(for/en)	#sent	#word(for/en)	#sent	#word(for/en)
da:Danish	855k	9.40/9.54M	1282k	25.8/27.4M	22k	0.177/0.204M	84k	0.824/0.835M
de:German	798k	8.50/8.85M	1235k	25.1/26.4M	68k	0.510/0.591M	88k	0.876/0.878M
el:Greek	820k	9.63/8.87M	710k	15.0/15.1M	220k	1.56/2.17M	63k	0.536/0.489M
es:Spanish	806k	9.62/8.65M	1289k	27.8/26.9M	495k	3.82/4.41M	101k	1.06/0.916M
fi:Finnish	851k	8.29/9.64M	1326k	20.7/28.6M	102k	0.586/0.906M	40k	0.292/0.330M
fr:French	778k	9.51/8.21M	1267k	28.4/25.9M	207k	1.72/1.86M	94k	1.138/0.805M
it:Italian	798k	9.46/8.68M	1214k	26.3/26.4M	16k	0.113/0.133M	98k	1.025/0.929M
nl:Dutch	831k	9.41/9.27M	1266k	26.8/26.7M	290k	2.14/2.58M	54k	0.533/0.538M
pt:Portuguese	831k	10.0/9.00M	1290k	27.7/27.2M	341k	2.46/2.93M	111k	1.214/1.038M
sv:Swedish	851k	9.04/9.59M	1248k	24.2/26.7M	320k	2.30/2.84M	104k	1.038/1.064M

Table 3: Training Bitext Statistics (#sent = number of sentences, #word = number of words on foreign and English side).

	in	out	in+out	i+o(A)	in+ir	in+ir(A)
da	**64.27**	38.81	63.82	62.97	63.20	63.33
de	**53.60**	31.53	53.13	53.20	**53.61**	52.83
el	**58.93**	26.22	58.27	58.48	**58.91**	**58.73**
es	60.96	43.87	**61.70**	60.60	61.02	60.98
fi	**52.40**	25.96	**52.46**	51.86	52.19	52.25
fr	**60.80**	45.44	60.59	60.24	60.63	**60.88**
it	61.61	44.14	62.16	61.81	61.51	**62.37**
nl	59.46	40.98	**59.74**	59.25	59.40	59.52
pt	**60.53**	44.27	59.69	59.87	59.72	60.31
sv	**60.89**	39.80	**60.86**	**60.90**	**61.01**	60.20
$\star$	7	0	4	1	3	3
$\lozenge$	in>{i+o(A),in+ir}, {in+ir(A),in+ir}>i+o(A)					

Table 7: EMEA task: test BLEU

	in	out	in+out	i+o(A)	in+ir	in+ir(A)
da	45.97	32.47	44.53	46.21	46.09	**46.65**
de	34.43	25.72	34.33	**35.50**	**35.31**	34.43
el	38.33	15.51	37.94	38.16	**39.04**	38.69
es	39.88	21.46	38.29	39.69	39.17	**40.24**
fi	18.00	11.87	**19.79**	17.19	18.29	17.40
fr	31.06	18.19	30.47	**31.83**	30.46	30.23
it	42.69	27.08	41.14	42.75	42.22	**43.29**
nl	34.63	25.60	**35.82**	**35.74**	**35.76**	34.21
pt	44.20	21.46	43.08	**45.25**	43.94	43.41
sv	**49.87**	28.53	48.54	49.73	**49.87**	**49.74**
$\star$	1	0	2	4	4	4
$\lozenge$	in+ir>in+ir(A), {in,i+o(A),in+ir(A),in+ir}>i+o					

Table 8: KDE task: test BLEU

features 4,5 (30% of weight), and those negatively correlated are features 2,3 (60% of weight). Thus, the more retrieved words, the more often **IN+IR(A)** outperforms; the larger the base OOV rate or improvement in OOV rate from IR data, the more often **IN+IR** outperforms.[5]

The one thing we can say with relative confidence is that large OOV rate is often predictive of **IN+OUT** and **IN+IR** giving improvements. We performed a data reduction experiment on EMEA to corroborate this. Out-of-domain data rarely helps in the original EMEA task, but as the in-domain data is reduced, word alignment and phrase coverage problems arise. Figure 2 shows that there is a consistent improvement for all language pairs in the case of 1/8 in-domain data (same is true for 1/4 data). We observe a strong correlation between lower OOV rate (as a result of out-of-domain phrases) and higher BLEU score (r=.9). On average, **IN+OUT** gives 2.54 BLEU improvement, **IN+IR** gives 1.8 improvement, and **IN+OUT(A)** gives 0.65 improvement. Thus, for the high OOV scenario, it appears that phrase coverage is a more important factor in improving BLEU.

[5]We also performed similar analyses for other pairs of systems, but the differences between SVM weights were not significant enough for us to make any conclusions.

5. Related Work

Prior work in translation model adaptation can be categorized by where in the pipeline out-of-domain data is applied to. For word alignment adaptation, [11, 12] explored ways to interpolate word alignment models trained independently on different domains. For phrase extraction, several works have explored using additional (possibly monolingual) data to improve coverage and reduce OOV rate. For example, [13, 14] finds paraphrase from monolingual corpora, and [15] translates the out-of-domain monolingual corpora to generate synthetic training data. Recent works using instance weighting techniques have also shown that out-of-domain data can improve phrase table scores [16, 17, 18]. The fact that all these methods give improvements is evidence that out-of-domain data can be beneficial to different parts of the training pipeline.

There are also works that do not target a specific step in the training pipeline. The most straightforward method is to directly combine in-domain and out-of-domain data, and run the combined dataset through the entire training pipeline (**IN+OUT**). A variant (**IN+IR**) is to add a subset by information retrieval methods [3, 4]. Alternatively, another approach

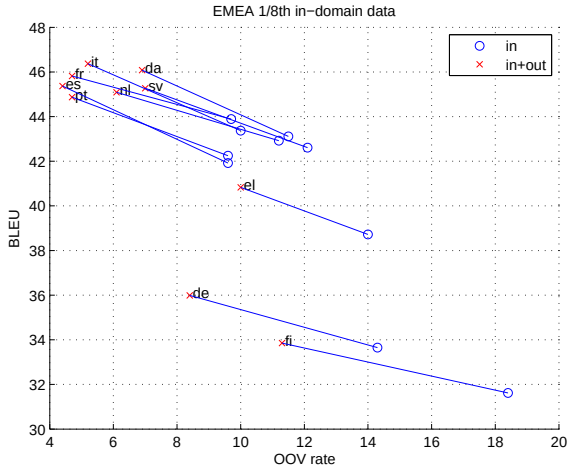

Figure 2: Data reduction experiment on 1/8th EMEA data.

is to run in-domain and out-of-domain data each separately through the training pipeline, and combine the final model by a log-linear model [19] or dynamic interpolation [20, 21]. These methods can be considered as combining data either before or after the training pipeline. While these methods are often effective, our experiments (e.g. **IN+IR(A)**) suggest it is also worthwhile to consider combining data for a part of the pipeline.

6. Conclusions

This paper empirically analyzed *where and when* out-of-domain data might help in the training pipeline. We confirmed our hypothesis that **out-of-domain data has different effects on the word alignment and phrase extraction steps**. This suggests a framework for designing new domain adaption techniques, where out-of-domain data is inserted at one step but excluded in another.

Our IWSLT experiments show that out-of-domain data improves word alignment by reducing the lexical translation ambiguity. Our experiments on EMEA, Europarl, KDE, and OpenSubtitles show that the results can be more varied. Based on these results, we make the following humble suggestions for practitioners of SMT domain adaptation:

1. Investigate various ways to combine out-of-domain data in the training pipeline, and perform model selection to obtain optimal results.
2. Quantify the improvements by analysis techniques (such as those presented in Sections 3.3) and use it to motivate the design of the next domain adaptation technique. For instance, our analyses suggest that word alignment adaptation that focuses only on in-domain words is a promising avenue for future research.

7. Acknowledgments

K.D. would like to thank Hal Daume for insightful discussions that provided motivation for part of this work.

8. References

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *HLT*, 2003.

[2] F. Och and H. Ney, "The alignment template approach to statistical MT," *Computational Linguistics*, 2004.

[3] S. Hildebrand, M. Eck, and S. Vogel, "Adaptation of translation model for SMT based on IR," *EAMT*, 2005.

[4] S. Huet, J. Bourdaillet, A. Patry, and P. Langlais, "The RALI MT system for WMT 2010," in *WMT*, 2010.

[5] Z. Li, et. al., "Joshua: an open source toolkit for parsing-based machine translation," in *WMT*, 2009.

[6] A. Fraser and D. Marcu, "Measuring word alignment quality for SMT," *Computational Linguistics*, 2009.

[7] N. Ayan and B. Dorr, "Going beyond AER: an extensive analysis of word alignments and their impact on MT," in *ACL*, 2006.

[8] F. Guzman, Q. Gao, and S. Vogel, "Reassessment of the role of phrase extraction in PBSMT," in *MT Summit XII*, 2009.

[9] P. Koehn, "Europarl: a parallel corpus for statistical machine translation," in *MT Summmit*, 2005.

[10] J. Tiedemann, "News from OPUS - collection of multilingual parallel corpora," in *RANLP*, 2009.

[11] H. Wu, H. Wang, and Z. Liu, "Alignment model adaptation for domain specific alignment," in *ACL*, 2005.

[12] J. Civera and A. Juan, "Domain adaptation in SMT with mixture modelling," in *WMT*, 2007.

[13] Y. Marton, C. Callison-Burch, and P. Resnik, "Improved statistical machine translation using monolingually-derived paraphrases," in *EMNLP*, 2009.

[14] M. Snover, B. Dorr, and R. Schwartz, "Language and translation model adaptation using comparable corpora," in *EMNLP*, 2008.

[15] N. Bertoldi and M. Federico, "Domain adaptation for SMT with monolingual resources," in *WMT*, 2009.

[16] S. Matsoukas, A.-V. I. Rosti, and B. Zhang, "Discriminative corpus weight estimation for MT," in *EMNLP*, 2009.

[17] G. Foster, C. Goutte, and R. Kuhn, "Discriminative instance weighting for domain adaptation in SMT," in *EMNLP*, 2010.

[18] K. Shah, L. Barrault, and H. Schwenk, "Translation model adaptation by resampling," in *WMT*, 2010.

[19] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical MT," in *WMT*, 2007.

[20] A. Finch and E. Sumita, "Dynamic model interpolation for statistical machine translation," in *WMT*, 2008.

[21] Y. Lü, J. Huang, and Q. Liu, "Improving statistical machine translation performance by training data selection and optimization," in *EMNLP-CoNLL*, 2007.