Domain adaptation problem in statistical machine translation systems

MARA CHINEA-RIOS [a], GERMÁN SANCHIS-TRILLES [b] and FRANCISCO CASACUBERTA [a]

[a] *Pattern Recognition and Human Language Technology Research Center, Valencia, Spain. e-mail: {machirio,fcn}@prhlt.upv.es*
[b] *Sciling, Valencia, Spain. {gsanchis}@sciling.com*

Abstract. Globalization suddenly brings many people from different country to interact with each other, requiring them to be able to speak several languages. Human translators are slow and expensive, we find the necessity of developing machine translators to automatize the task. Several approaches of Machine translation have been develop by the researchers. In this work, we use the Statistical Machine Translation approach. Statistical Machine Translation systems perform poorly when applied on new domains. The domain adaptation problem has recently gained interest in Statistical Machine Translation. The basic idea is to improve the performance of the system trained and tuned with different domain than the one to be translated. This article studies different paradigms of domain adaptation. The results report improvements compared with a system trained only with in-domain data and trained with all the available data.

Keywords. statistical machine translation, domain adaptation, data selection, data combination, phrase tables

Introduction

Natural Language Processing (NLP) is a research field of artificial intelligence and linguistics that is gaining importance with the up-rise of computerised communication technologies.

Machine Translation (MT) is a specific sub-field of NLP, and studies the way in which automatic systems should be developed so that they are able to translate a certain sentence in a source language into a sentence in a given target language, such that source and target sentences preserve the exact same meaning, while being both well-formed sentences in their respective languages.

Several approaches have been used for this task with a different level of success. The first and most intuitive way is the *word-for-word* translation, the principal problem with the word order or the context make difficult to understand the meaning of the translated sentence. Other approach is Rule-based Machine Translation (RBMT). The first RBMT systems were developed in the early 1970s and were the first commercial machine translation systems. The RBMT systems are based on linguistic rules that allow the words to be put in different places and to have different meaning depending on context.

Bilingual corpora are precious resources in computational linguistics and they are the possibility of performing another kind of automatic translation methods. This is the case of the Statistical Machine Translation (SMT), transformed the state of the art in MT completely. The goal is to create mathematical models that can describe the translation process accurately and then, estimate the translation and ordering probabilities automatically using the training corpus.

SMT have a great potential but they are not able to provide ready to use translations in real-world applications and many researchers are working on it. SMT relies heavily on the availability of such bilingual corpora. Usually, bilingual corpora are used to estimate the parameters of the translation model. Unfortunately, we do not have parallel data in all domains. For this reason, the translation quality gets worse when we do not have enough training data for the specific domain we need to tackle in our test set.

The domain adaptation problem is very common in SMT, where the objective is to improve the performance of systems trained and tuned on out-of-domain corpus by using very limited amounts of in-domain corpus.

The main contribution of this paper is:

- We compare two different domain adaptation paradigms. To the best of our knowledge, such study does not exist in the literature.

This paper is structured as follows. Section 2 presents statistical machine translation formulation. Section 3 summarises the related work domain adaptation paradigms. In Section 4, experimental results are reported. Finally, conclusions and future work are presented in Section 5.

1. Statistical machine translation

The grounds of modern SMT were established in [1] where the problem of machine translation was defined as the problem of translating a source sentence $\mathbf{x} = x_1 ... x_J$ into a target sentence $\mathbf{y} = y_1 ... y_I$. Basically, given a sentence $\mathbf{x}$ from a source language, we want to find the most likely sentence $\mathbf{y}$ from a target language that corresponds to its translation:

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} Pr(\mathbf{y}|\mathbf{x}) \tag{1}$$

Using this expression, it seems that we only have to look for the target sentence $\mathbf{y}$ with the highest probability, given a source sentence $\mathbf{x}$. But it will require to estimate in advance the probabilities for any pair of arbitrary sentence pairs ($\mathbf{y}$, $\mathbf{x}$) and that is impossible, because of the lack of such huge corpora. Instead of it, using the *Bayes* rule:

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} \frac{Pr(\mathbf{y}) \cdot Pr(\mathbf{x}|\mathbf{y})}{Pr(\mathbf{x})} \tag{2}$$

Noticing that the term $Pr(\mathbf{x})$ does not influence on the search for the arguments maximizing the fraction, we obtain:

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} Pr(\mathbf{y}) \cdot Pr(\mathbf{x}|\mathbf{y}) \tag{3}$$

Here, $Pr(\mathbf{x}|\mathbf{y})$ is a term representing the translation model, based on correspondences between the two languages, and $Pr(\mathbf{y})$ stands for the language model, usually based on n-grams.

Recently, the direct modelling of the posterior probability $Pr(\mathbf{x}|\mathbf{y})$ has been widely adopted. Different authors [2,3] propose the use of the so-called log-linear models, where the decision rule is given by the expression

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\text{argmax}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{x}, \mathbf{y}) \tag{4}$$

where λ_m is the weight assigned to $h_m(\mathbf{x}, \mathbf{y})$ and $h_m(\mathbf{x}, \mathbf{y})$ is a score function representing an important feature for the translation of $\mathbf{x}$ into $\mathbf{y}$, as for example the language model of the target language, a reordering model, or several translation models. The weights λ_m are normally optimised with the use of a development set. The most popular approach for adjusting λ_m is the one proposed in [4], commonly referred to as Minimum Error Rate Training (MERT). This algorithm implements a coordinate-wise global optimisation.

2. Domain adaptation in SMT

As it was anticipated during the introduction, translating text that belongs to a different domain than the bilingual corpora used for training and tuning leads to lower translation quality. This gives rise to the very common problem of domain adaptation, where the objective is to improve the performance of the system on the specific domain being tackled.

The standard consists in training SMT systems with all the available data. It is assumed that the more data used to train the system, the better. This assumption is correct if all the data belongs to the same domain. However, this is not the case in the problems tackled by most of the SMT systems. In fact, most SMT systems are designed to translate specific text, such as user manuals or medical prospects.

In the next section, we briefly review the state of art of domain adaptation, divided into two different paradigms: data selection and data combination.

We will refer to the pool of sentences available as *out-of-domain* corpus because we assume that it belongs to a different domain than the one to be translated. Similarly, we refer to the corpus of the domain of the text to translate as *in-domain* corpus.

2.1. Data selection

Data selection (DS) aims to select the best subset of bilingual sentences from an available out-to-domain. By doing so, we pretend to improve the state of the art in terms of translation quality obtained and computational requirements, without using the complete pool sentences.

State-of-the-art DS approaches rely on the idea of choosing those sentence pairs in the out-of-domain training corpus that are in some way similar to an in-domain training corpus in terms of some different metrics.

The simplest instance of this problem can be found in language modelling, where perplexity-based selection methods have been used [5]. Here, out-of-domain sentences are ranked by their perplexity score. Another perplexity-based approach is presented in [6], where cross-entropy difference is used as a ranking function rather than just perplexity, in order to account for normalization. We apply this criterion for the task of selecting training data for SMT systems

Different works use perplexity-related DS strategies [7,8,9,10,11]. In these papers, the authors report good results when using the strategy presented in [6], and such strategy has become a de-facto standard in the SMT research community. In [8] the authors describe the *XenC* open source toolkit for data selection. *XenC* uses the two strategies described in [5] and [6]. The best results were obtained using difference in cross-entropies. In our experiments with cross-entropy, we will be using *XenC*.

Two different approaches are presented in [12]: one based on approximating the probability of an in-domain corpus and another one based on infrequent n-gram recovery. The technique approximating the probability relies on preserving the probability distribution of the task domain by wisely selecting the bilingual pairs to be used. Hence, it is mandatory to exclude sentences from the pool that distort the actual probability. The technique based in infrequent n-gram recovery consists in increasing the information of the in-domain corpus by adding evidence for those n-grams that have been seldom observed in the in-domain corpus. This evidence is obtained by selecting sentences from the out-of-domain corpus. The n-grams that have never been seen or have been seen just a few times are called *infrequent n-grams*. The best results were obtained with infrequent n-grams recovery, achieving an improvement of 1 BLEU point.

Other works have applied information retrieval methods for DS [13], in order to produce different sub-models which are then weighted. In that work, authors define the baseline as the result obtained by training only with the corpus that shares the same domain with the test. Afterwards, they claim that they are able to improve the baseline translation quality by adding new sentences retrieved with their method. However, they do not compare their technique with a model trained with all the corpora available.

More recently, [14] utilised neural language models to perform data selection, reporting substantial gains over conventional n-gram language model-based data selection across multiple language pairs.

2.2. *Data combination for phrase tables*

Studies in data selection techniques have typically focused on how to select the best subset of the out-of-domain corpus so as to concatenate it with the in-domain corpus, and then such concatenation is used for training the final SMT system. In this section, we present different approaches present in the literature for combining the in-domain and out-of-domain models, with the purpose of using such approaches for combining the in-domain model with the model trained on the selected data.

In [15] a mixture model approach is proposed. The authors explored different choices: linear and log-linear mixtures. The result show improvements by the linear and log-linear mixtures over a baseline trained with all training data.

In [16] the authors adapted a phrase-based SMT system to the new domains by integrating it with language and translation models. Phrase-pairs are here scored with four translation probabilities and four reordering probabilities, thus resulting in a significantly larger set of feature weights to be trained.

In [17] the authors presented fill-up method and they compare his method with standard interpolation methods. Fill-up method is applied after a standard phrase-based SMT training process and just before weight optimization. Fill-up effectively exploits background knowledge to improve model coverage, while preserving the more reliable information coming from the in-domain corpus. First, we separate translation models are built from in-domain corpus and out-of-domain corpus. In fill-up method out-of-domain table is merged with in-domain table by adding only new phase pairs that no appear in the in-domain table.

In [7] the authors used three methods based in cross-entropy for extracting a pseudo in-domain corpus. This pseudo in-domain corpus is used to train a small domain-adapted SMT system. The authors combined the small domain-adapted translation model with the true in-domain translation model via linear and log-linear mixtures. In the reported experiments, both mixture methods outperformed the in-domain and general baselines.

Finally, in [18] a corpus identifier is introduced to distinguish the parallel in-domain corpus from the out-of-domain corpus in a factored translation model. Each target word is assigned an id tag corresponding to the part of the corpus it belongs to. Three additional translation model features are introduced to compute the probability of the corpus id tags being generated given the source phrase, as well as the source and target phrase probabilities, given the corpus id tags. The incorporation of corpus id tags promotes the preference of phrase pairs from a specific domain.

2.3. Experimental set-up

We evaluated empirically the domain adaptation methods described in the previous section. For the out-of-domain corpus, we used the English-French parallel text from release v7 of the Europarl [1] corpus [19]. The Europarl corpus is composed of translations of the proceedings of the European parliament. As in-domain data, we used the EMEA[2] corpus [20] is available in 22 languages and contains documents from the European Medicines Agency. We evaluated our work on the Khresmoi Summary 2014[3] test set. The main figures of the corpora used are shown in Tables 1 and 2.

		EN	FR
Europarl	$\|S\|$	2.2M	
	$\|W\|$	50.2M	52.5M
	$\|V\|$	157k	215k

Table 1. Europarl corpus main figures. k denotes thousands of elements, M denotes million of elements, $|S|$ stands for number of sentences, $|W|$ stands for number of words (tokens) and $|V|$ for vocabulary size (types).

All experiments were carried out using the open-source SMT toolkit Moses version phrase-based [21]. The language model used was a 5-gram, standard in SMT research, with modified Kneser-Ney smoothing [22], built with the SRILM toolkit [23]. The phrase table was generated by means of symmetrised word alignments obtained with GIZA++ [24]. The decoder features a statistical log-linear model including a phrase-based translation model, a language model, a distortion model and word and phrase penalties. The log-lineal combination weights in Equation 4 were optimized using MERT (minimum

[1] `www.statmt.org/europarl/`

[2] `www.opus.lingfil.uu.se/EMEA.php`

[3] `www.statmt.org/wmt14/medical-task/`

	EMEA-Domain		Medical-Test		Medical-Mert	
	EN	FR	EN	FR	EN	FR
$\|S\|$	1.0M		1000		501	
$\|P\|$	12.1M	14.1M	21.4k	26.9k	9850	11.6k
$\|V\|$	98.1k	112k	1.8k	1.9k	979	1.0k

Table 2. Medical main figures. EMEA-Domain is the in-domain corpus, Medical-Test is the evaluation data and Medical-Mert is development set. M denotes millions of elements and k thousands of elements, $|S|$ stands for number of sentences, $|W|$ for number of words (tokens) and $|V|$ for vocabulary size (types).

error rate training) [4] on the Mediacal-Mert data, which was the development set used in the 2014 WMT evaluation.

We compared the selection methods with two baseline systems. The first one was obtained by training the SMT system with EMEA-Domain data. We will refer to this setup with the name of `baseline-emea`. A second baseline experiment has been carried out with the concatenation of the Europarl corpus and EMEA training data. We will refer to this setup with the name of `baseline-all`.

Evaluation in SMT is a very controversial issue. Human evaluation is way too costly for experimentation purposes. This leads to the wide-spread use of automatic evaluation metrics that are very cheap to use. In this work, SMT output will be evaluated by means of BLEU [25] and TER [26], which are two of the most popular evaluation metrics employed in SMT.

- BLEU (Bilingual Evaluation Understudy) score: This score measures the precision of uni-grams, bigrams, trigrams, and four-grams with respect to a set of reference translations, with a penalty for too short sentences [25]. BLEU is not an error rate, i.e. the higher the BLEU score, the better. BLEU will be reported as a percentage, ranging from 0 to 100.
- TER (Translation Edit Rate): Translation Error Rate TER [26] is an error metric for MT that measures the number of edits required to change a system output into one of the references. TER will also be reported as a percentage.

2.4. Results for data selection methods

In this section, we present the experimental results obtained with two strategies for data selection methods (Cross-entropy method and infrequent n-grams recovery) presented in Section 2.1.

Table 3 shows the principal results obtained cross-entropy selection and infrequent n-grams recovery. Several conclusion can be drawn:

- The translation quality provided by the both methods is better in term of BLEU and TER than the results achieved with the system Baseline-emea and Baseline-all.
- Lastly, it is also worth noting that the results obtained with the cross-entropy selection are slightly worse than the ones obtained with infrequent n-grams recovery in all the set-ups analysed, even though more sentences are considered when using cross-entropy.

Domain	Strategy	BLEU	TER	$\|S\|$
Medical	Baseline-emea	28.5	53.2	1.0M
	Baseline-all	29.4	53.6	1.0M + 1.4M
	Cross-entropy	29.7	52.6	1.0M + 200K
	Infrequent n-grams	30.2	51.6	1.0M + 44K

Table 3. Summary of the best results obtained with each methods DS. $|S|$ for number of sentences, which are given in terms of the in-domain corpus size, and (+) the number of sentence selected.

2.5. Results for data combination for phrase tables

In this section, we present the experimental results obtained with two data combination strategies for phrase tables (Fill-up and linear interpolation methods) presented in Section 2.2.

Table 4 shows the principal results and the two baseline systems evaluated on the Medical-Test set. Several conclusion can be drawn:

- The translation quality provided by the Fill-up method is large better in term of BLEU and TER than the results achieved with the system Baseline-emea and Baseline-all.
- Lastly, it is also worth noting that the results obtained with the linear interpolation method are slightly worse than the ones obtained with fill-up method.

Domain	Strategy	BLEU	TER
Medical	Baseline-emea	28.5	53.2
	Baseline-all	29.4	53.6
	Interpolation	29.3	53.9
	Fill-up	30.0	53.2

Table 4. Summary of the best results obtained with each data combination methods for phrase-based SMT. $|S|$ for number of sentences, which are given in terms of the in-domain corpus size, and (+) the number of sentence selected.

2.6. Example Translations

Translation example are shown in Table 5. In the first example, both the infrequent n-gram selection and baseline systems are able to obtain the character $\%$ as appears in the reference. This in not only casual, since, by ensuring coverage for the infrequent n-grams only up to a certain t, we avoid distorting the specificities of the in-domain data. All the systems present the same lexical choice error with word (*développer*). However, this is so because this is the most likely translation in our data, both in-domain and out-of-domain.

In the second example, all the systems present the same lexical choice error with word (*d′ autres*). However, this is so because this is the most likely translation in our data, both in-domain and out-of-domain.

Src	about 5 percent of people with ulcerative colitis develop colon cancer .
Bsl	environ 5 % des personnes avec colite ulcÃl’reuse *de développer* un cancer du colon .
All	environ 5 *pour cent* des personnes avec colite ulcéreuse *développer* un cancer du colon .
Infr	environ 5 % des personnes avec colite ulcéreuse de *développer* un cancer du colon .
Entr	environ 5 *pour cent* des personnes avec colite ulcéreuse de *développer* un cancer du colon .
Fill-up	environ 5 *pour cent* des personnes avec colite ulcéreuse de *développer* un cancer du colon .
Intp	environ 5 *pour cent* des personnes avec colite ulcéreuse de *développer* un cancer du colon .
Ref	environ 5 % des personnes souffrant de colite ulcéreuse sont atteintes de cancer du côlon.
Src	other patients with any of the above symptoms should consult their doctor .
Bsl	autres les patients avec un des symptômes ci-dessus doivent consulter leur médecin .
All	autres les patients avec un des symptômes ci-dessus doivent consulter leur médecin .
Infr	autres les patients avec un des symptômes ci-dessus doivent consulter leur médecin .
Entr	autres les patients avec un des symptômes ci-dessus doivent consulter leur médecin .
Fill-up	autres les patients avec un des symptômes ci-dessus doivent consulter leur médecin .
Intp	autres les patients avec un des symptômes ci-dessus doivent consulter leur médecin .
Ref	d ′ autres patients avec un des symptômes ci-dessus devraient consulter leur médecin .

Table 5. Example of one translations for each of the SMT systems built: Src (source sentence), Bsl (baseline), All (all the data available), Infr (Infrequent n-grams), Entr (Cross-entropy), Fill-up (Fill-up method), Intp (Linear interpolation method) and Ref (reference).

3. Conclusion and future work

Domain adaptation has been receiving an increasing amount of interest within the SMT research community. There are a lot of domain adaptation methods. In this work, we study two different domain adaptation paradigm. In this work, we perform a comparison of four techniques (Two data selection and two data combination methods). The results obtained are very similar, although the best results were obtained by the infrequent n-grams recovery using only 4% of the out-to-domain corpus.

In future work, we intend to combine the two paradigms proposed and will develop new experiments with bigger and more diverse data sets.

Acknowledgements

The research leading to these results has received funding from the Generalitat Valenciana under grant Prometeo/2009/014 and the FPI grant by Universitat Politècnica de València.

References

[1] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” *Computational linguistics*, vol. 19, no. 2, pp. 263–311, (1993).

[2] K. A. Papineni, S. Roukos, and R. T. Ward, “Maximum likelihood and discriminative training of direct translation models,” in *Proceedings of International Conference on Acoustics, Speech and Signal*, pp. 189–192, (1998).

[3] F. J. Och and H. Ney, “Discriminative training and maximum entropy models for statistical machine translation,” in *Proceedings of Association for Computational Linguistics*, pp. 295–302, (2002).

[4] F. J. Och, “Minimum error rate training in statistical machine translation,” in *Proceedings of Association for Computational Linguistics*, pp. 160–167, (2003).

[5] J. Gao, J. Goodman, M. Li, and K.-F. Lee, “Toward a unified approach to statistical language modeling for chinese,” *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 1, no. 1, pp. 3–33, 2002.

[6] R. C. Moore and W. Lewis, “Intelligent selection of language model training data,” in *Proceedings of Association for Computational Linguistics*, pp. 220–224, (2010).

[7] A. Axelrod, X. He, and J. Gao, “Domain adaptation via pseudo in-domain data selection,” in *Proceedings of Empirical Methods in Natural Language Processing*, pp. 355–362, (2011).

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

[9] H. Schwenk, A. Rousseau, and M. Attik, “Large, pruned or continuous space language models on a gpu for statistical machine translation,” in *Proceedings of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies*, pp. 11–19, (2012).

[10] R. Sennrich, “Perplexity minimization for translation model domain adaptation in statistical machine translation,” in *Proceedings of European Chapter of the Association for Computational Linguistics*, pp. 539–549, (2012).

[11] S. Mansour, J. Wuebker, and H. Ney, “Combining translation and language model scoring for domain-specific data filtering.,” in *Proceedings of International Workshop on Spoken Language Translation*, pp. 222–229, (2011).

[12] G. Gascó, M.-A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta, “Does more data always yield better translations?,” in *Proceedings of European Chapter of the Association for Computational Linguistics*, pp. 152–161, (2012).

[13] Y. Lü, J. Huang, and Q. Liu, “Improving statistical machine translation performance by training data selection and optimization.,” in *Proceedings of Empirical Methods in Natural Language Processing-CoNLL*, pp. 343–350, (2007).

[14] K. Duh, G. Neubig, K. Sudoh, and H. Tsukada, “Adaptation data selection using neural language models: Experiments in machine translation.,” in *Proceedings of Association for Computational Linguistics*, pp. 678–683, 2013.

[15] G. Foster and R. Kuhn, “Mixture-model adaptation for smt,” in *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 128–135, (2007).

[16] P. Koehn and J. Schroeder, “Experiments in domain adaptation for statistical machine translation,” in *Proceedings of Workshop on Statistical Machine Translation*, pp. 224–227, (2007).

[17] A. Bisazza, N. Ruiz, M. Federico, and F.-F. B. Kessler, “Fill-up versus interpolation methods for phrase-based smt adaptation.,” in *Proceedings of International Workshop on Spoken Language Translation*, pp. 136–143, (2011).

[18] J. Niehues and A. Waibel, “Domain adaptation in statistical machine translation using factored translation models,” in *Proceedings of European Association for Machine Translation*, (2010).

[19] P. Koehn, “Europarl: A parallel corpus for statistical machine translation,” in *Proceedings of MT summit*, pp. 79–86, (2005).

[20] J. Tiedemann, “News from opus-a collection of multilingual parallel corpora with tools and interfaces,” in *Proceedings of Recent advances in natural language*, pp. 237–248, (2009).

[21] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: open source toolkit for statistical machine translation,” in *Proceedings of Association for Computational Linguistics*, pp. 177–180, (2007).

[22] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 181–184, (1995).

[23] A. Stolcke, “Srilm-an extensible language modeling toolkit,” in *Proceedings of International Conference on Spoken Language Processing*, (2002).

[24] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational linguistics*, vol. 29, pp. 19–51, (2003).

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of Association for Computational Linguistics*, pp. 311–318, (2002).

[26] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A study of translation edit rate with targeted human annotation,” in *Proceedings of Association for Machine Translation in the Americas*, pp. 223–231, (2006).