Domain Adaptation for Social Localisation-based SMT: A Case Study Using the Trommons Platform

Jinhua Du jdu@computing.dcu.ie
Andy Way away@computing.dcu.ie
Zhengwei Qiu zhengwei.qiu@computing.dcu.ie
ADAPT Centre, School of Computing, Dublin City University, Ireland

Asanka Wasala Asanka.Wasala@ul.ie
Reinhard Schaler reinhard.schaler@ul.ie
Localisation Research Centre, University of Limerick, Ireland

Abstract

Social localisation is a kind of community action, which matches communities and the content they need, and supports their localisation efforts. The goal of social localisation-based statistical machine translation (SL-SMT) is to support and bridge global communities exchanging any type of digital content across different languages and cultures. Trommons is an open platform maintained by The Rosetta Foundation to connect non-profit translation projects and organisations with the skills and interests of volunteer translators, where they can translate, post-edit or proofread different types of documents. Using Trommons as the experimental platform, this paper focuses on domain adaptation techniques to augment SL-SMT to facilitate translators/post-editors. Specifically, the Cross Entropy Difference algorithm is used to adapt Europarl data to the social localisation data. Experimental results on English–Spanish show that the domain adaptation techniques can significantly improve translation performance by 6.82 absolute BLEU points and 5.99 absolute TER points compared to the baseline.

1 Introduction

The concept of social localisation was proposed In Schäler (2011), which is based on the fact that currently large communities with language skills are ready to support good causes, and large amounts of content are accessible to communities that should be translated, but is not being done currently. Accordingly, social localisation aims to match communities and slice content, and support the localisation efforts.[1]

The main objective of social localisation is the promotion of a demand– rather than a supply–driven approach to localisation. It is based on the recognition that it is no longer exclusively the fact that the corporations control the global conversation, but rather communities. Social localisation supports user-driven and needs-based localisation scenarios, in contrast to mainstream localisation, driven primarily by short-term financial return-on-investment considerations.

Trommons[2] - short for Translation Commons - is an open platform maintained by The Rosetta Foundation[3] to connect non-profit translation projects and organisations with the skills

[1] https://en.wikipedia.org/wiki/Social_localisation
[2] http://trommons.org/
[3] http://www.therosettafoundation.org/

and interests of volunteer translators. The idea behind Trommons is to match translation/post-editors/proofreading tasks published by various NGOs with volunteer translators/users registered in Trommons.

It is well-known that SMT systems have been widely deployed into the translator's workflow in the localisation and translation industry to improve productivity, which is also named as post-editing SMT (PE-SMT) (Guerberof, 2009; Plitt and Masselot, 2010; Carl et al., 2011; O'Brien, 2011; Zhechev, 2012; Guerberof, 2013). In Trommons, many language pairs lack high-level translators, or translation jobs take quite a long time to be claimed and finished, so a PE-SMT system for such language pairs or translation projects are a practical solution in which human effort could be significantly reduced and high-quality translations could be achieved.

However, data is a big problem for building high-quality PE-SMT systems in Trommons. In this paper, we use domain adaptation techniques to acquire social localisation domain related data to augment PE-SMT systems. Specifically, a Cross Entropy Difference (CED) method is used to select different scales of data from the Europarl data sets. Experiments conducted on English-to-Spanish show that the domain adaptation method can improve translation quality by 6.82 absolute (20.98 relative) BLEU points (Papineni et al., 2002) and 5.99 absolute (11.26 relative) TER points (Snover et al., 2006) compared to the baseline.

2 Related Work

Domain adaptation is a popular but difficult research question in SMT. It is well-known that SMT performance is heavily dependent on the training data and the development set (devset) as well as the estimated model parameters which can best reflect the characteristics of the training data. Therefore, if the characteristics of the test data are substantially different from those of the model parameters, system performance drops significantly. In this case, the domain adaptation can be used to adapt an SMT system to the out-of-domain (OOD) test data in order to improve translation performance.

For social localisation data in Trommons, the currently available in-domain (ID) data for each language pair is far from sufficient to build up a high-quality SMT system (cf. Section 5.2). Therefore, we have to use a data selection algorithm to select ID data from the OOD data in order to augment the SL-SMT.

Lv et al. (2007) use information-retrieval techniques in a transductive learning framework to increase the count of important ID training instances, which results in phrase-pair weights being favourable to the devset. Bicici and Yuret (2011) employ a Feature Decay Algorithm (FDA) to increase the variety of the training set by devaluing features that have already been seen from a training set, and experimental results show significant improvements compared to the baseline.

There has been increasing interest in applying the CED method to the problem of SMT data selection (Moore and Lewis, 2010; Axelrod et al., 2011; Haque et al., 2014). In this method, given an ID corpus I and a general corpus O, language models are built from both, and each sentence s in O is scored according to the entropy difference, as in Equation (1):

$$score(s) = H_I(s) - H_O(s) \quad (1)$$

where $H_I(s)$ is the entropy of s in ID corpus I, and $H_O(s)$ is the entropy of s in OOD corpus O. The $score(s)$ is to reflect how similar the sentence s is to the corpus I, and how different the sentence s is from the corpus O. That is, the lower the score given to a sentence, the more useful it is to train a system for the specific domain I.

Some other methods have been proposed to select ID data. For example, Banerjee et al. (2012) propose an approach to perform batch selection with the objective of maximizing SMT performance. Toral (2013) and Toral et al. (2014) use linguistic information such as lemmas,

named entity categories and part-of-speech tags to augment perplexity-based data selection and achieved improved results.

The CED method is more promising both in domain adaptation for SMT and data selection for active learning-based PE-SMT (Dara et al., 2014), so we choose it to verify its effectiveness in SL-SMT application.

3 System Description of Trommons

In Trommons, registered organisations/NGOs can create projects. Projects are identified by project IDs. A project can have a **source document** and a number of **tasks** associated with it. These tasks can be one of the following types: **translation, proofreading** or **segmentation**. All the files that belong to a certain project are stored in the project folder.

The system stores raw files in a MySQL database. The files can be in different formats (e.g. pdf, doc, xliff, ppt). Multiple versions of the files are stored in the file system, i.e. version 0 is normally the source file (v-0 folder), subsequent versions may denote the translation or proof-read version of the source file (e.g. v-1, v-2 folders). In the MySQL database, **Project table** contains the necessary information about the project; **Tasks table** includes different information about the tasks, including the language pairs, countries, organisations etc. Different task types are defined in the **TaskTypes table**, e.g., 3 indicates "Proofreading" and 2 indicates "Translation". The tasks can be in different states (claimed, completed etc.) which are defined in **TaskStatus** table. Using these tables a query can be used to retrieve the information, e.g. an organisation can be retrieved from the **Organisations table**.

Some projects may have been already completed, and the system allows archiving of such old projects. Data related to these projects can be retrieved from the **Archived Projects table** and the **Archived Tasks table**. Their associated metadata can be found in the **ArchivedProjectsMetadata table**. The **TaskFileVersions table** includes details of the latest version of the files (i.e. v-0, v-1 etc in the file system).

However, in Trommons, the parallel texts are not automatically maintained in the file system, so we have to extract and align source and target sentences from the raw files. In addition, information about the language pairs is not maintained in the file system; it can only be acquired from the database.

4 Data Statistics and Analysis

We export data from the current Trommons database, and examine the tables and data according to descriptions of the file system.

4.1 Data Statistics

The data statistics are listed in Table 1.

#Projects	#Tasks	#Organisations	#Languages	#Countries	#FileTypes
2,270	7,825	263	7,431	250	xls,doc,docx,zip,pdf,odt, dot,ppt,po,xlsx,xliff

Table 1: Statistics of the Trommons database

In Table 1, '#Projects' indicates the number of projects in the database, '#Tasks' is the total number in all projects, '#Organisations' is the number of different 'Organisations' registered in Trommons, '#Languages' indicates the number of different languages (including dialects, sign language etc.) that are used in projects, '#Countries' is the number of different countries where translators or organisations come from, and '#FileTypes' indicates different types of documents stored in the database.

From the data statistics, we can see that Trommons has successfully connected non-profit translation projects and organisations with thousands of volunteer translators worldwide.

4.2 Data Analysis

We query the 'completed translation' tasks from the database, and retrieve **1,990** tasks that belong to **964** projects, containing **23** languages in terms of the source language and **63** languages in terms of the target language, coming from **44** countries regarding the source language and **74** countries regarding the target language. In these 1,990 tasks, there are **101** language pairs in total.

The breakdown of **TOP-10** ranks in terms of language pairs, source language, target language and countries are shown in Table 2∼Table 4.

No.	source language	target language	#tasks	ratio(%)
1	English	Spanish	422	21.21
2	Spanish	English	266	13.37
3	English	French	152	7.64
4	French	English	105	5.28
5	English	Portuguese	94	4.72
6	English	Russian	87	4.37
7	English	Arabic	76	3.82
8	English	Chinese	71	3.57
9	Spanish	French	65	3.27
10	French	Spanish	59	2.96
total	–	–	**1,397**	**70.20**

Table 2: Top-10 language pairs in terms of tasks

In Table 2, we can see that 1) English-based language pairs have the largest proportion in terms of the source language; 2) Spanish-based language pairs have the largest proportion in terms of the target language; 3) the tasks of the top-10 language pairs account for 70.2% in all 1,990 tasks.

No.	source language-based tasks			target language-based tasks		
	language	#tasks	ratio(%)	language	#tasks	ratio(%)
1	English	1,280	64.32	English	499	25.08
2	Spanish	389	19.55	Spanish	499	25.08
3	French	167	8.39	French	224	11.26
4	Portuguese	29	1.46	Portuguese	116	5.83
5	German	27	1.36	Russian	88	4.42
6	Catalan	24	1.21	Arabic	79	3.97
7	Italian	18	0.90	Chinese	71	3.57
8	Hebrew	10	0.50	Italian	46	2.31
9	Chinese	9	0.45	Vietnamese	38	1.91
10	Swedish	7	0.35	Korean	37	1.86
total	–	**1,960**	**98.49**	–	**1,697**	**85.28**

Table 3: Top-10 languages in terms of source side and target side, respectively

In Table 3, we can see that 1) English-based tasks account for the largest proportion in terms of the source-side language; 2) English-based and Spanish-based tasks account for the

same proportion that is much greater than others in terms of the target language; 3) the tasks of the top-10 source-side languages and target-side languages account for 98.49% and 85.28%, respectively.

No.	countries based on source-side language			countries based on target-side language		
	country	#tasks	ratio(%)	country	#tasks	ratio(%)
1	United States	540	27.14	United States	377	18.94
2	United Kingdom	474	23.82	ANY	281	14.12
3	Spain	195	9.80	Spain	225	11.31
4	Ireland	156	7.84	United Kingdom	195	9.80
5	ANY	148	7.74	France	162	8.14
6	France	139	6.98	Ireland	95	4.77
7	Colombia	67	3.37	Peru	81	4.07
8	Mexico	52	2.61	Brazil	55	2.76
9	Brazil	29	1.46	Portugal	53	2.66
10	Congo	27	1.36	Russian Federation	53	2.66
total	–	**1,827**	**91.81**	–	**1,577**	**79.25**

Table 4: Top 10 countries in terms of source-side and target-side languages, respectively

In Table 4, 'ANY' indicates the country that was not specified when creating a task. We can see that 1) United States has the largest proportion of translation tasks in terms of both source-side language and target-side language; 2) Other countries such as United Kingdom, Spain, France and Ireland have a large proportion compared to the rest of the countries.

We can conclude from the statistics that in Trommons:

- English–Spanish is the most popular language pair, which also indicates that the number of translators of this language pair is the largest;
- English-oriented and Spanish-oriented tasks account for larger proportions than other languages;
- The United States, United Kingdom, Spain and France account for most of the tasks, but countries like Colombia, Mexico, Brazil, Peru account for big proportions as well because these countries use Spanish and Portuguese.

5 Experiment Preparation

Having examined what data is available, we now report on two experiments we carried out, namely the ID SMT experiment and domain adaptation SMT experiment. In the latter, the CED data selection method is used to acquire more ID data from the OOD data. The Experimental details are described in the following sections.

5.1 Experiment Plans and Considerations

We carried out the following steps in our experiments:

- Examination of the tasks that are "translation" type and which have been "completed", and recognition of how many language pairs and file types can be retrieved.
- Extraction of parallel data from the task files above.
- Sentence alignment and creating and evaluating sentence-level parallel data, and evaluate the parallel data.

- Creation of development set and test set.
- Development of SMT system and tuning of parameters.
- Evaluation of the SMT system.
- Selection of data from Europarl using CED method and building of domain–adaptive SL-SMT, and evaluation of its performance.

5.2 Data Pre-processing and Sentence Alignment

In Table 2, the English–Spanish pair has the biggest proportion of tasks among all language pairs, so we selected this for SMT system creation.

262 pairs of documents we retrieved for the English–Spanish pair, in which 231 documents were '.doc' or '.docx' files. These types of documents needed to be converted to '.txt' files for SMT system building.

We used the 'Hunalign' toolkit (Varga et al., 2005) to run the sentence alignment, and obtained 10,151 aligned sentence pairs. After filtering out the longer (>100) sentences, we ended up with 9,379 parallel sentences.

We then randomly selected out 100 sentence pairs in all 9,379 parallel sentences to evaluate the accuracy of aligned sentences by human, and we achieve the accuracy of **98%**.

6 SMT Experiments

Due to the insufficient amount of parallel data extracted from the Trommons tasks, domain adaptation techniques had to be used to acquire more ID parallel data to obtain reasonable translation performance.

In order to build the SMT system, we use Moses (Koehn et al., 2007) with default settings. For the data sets, we randomly extracted two sets of 500 sentence pairs from the 9,379 pairs of sentences as the devset and test set, respectively. The remaining 8,379 pairs of parallel data was used to build the initial ID SL-SMT system.

We used the CED method to select different scales of ID data from the Europarl data set. The statistics of the different adapted data sets are shown in Table 5.

system	#sentence	#word	
		English	Spanish
baseline0	8,379	171,593	185,852
baseline1	3,993,529	104,280,921	109,983,391
baseline2	3,993,529	104,280,921	109,983,391
DaM1	120,566	2,671,412	2,685,671
DaM2	205,593	5,171,210	5,500,986
DaM3	284,370	7,671,080	8,058,164
DaM4	363,103	10,170,859	10,645,559
DaM5	827,362	25,169,906	26,187,163
DaM6	982,405	30,169,652	31,366,487
DaM7	1,138,302	35,169,395	36,550,290
DaM8	1,295,919	40,169,132	41,745,438
DaM9	1,456,383	45,168,909	46,958,444
DaM10	1,620,025	50,168,648	52,184,793

Table 5: Statistics of different data sets

In Table 5, "baseline0" indicates that the data is the extracted parallel data from the Trommons platform. "baseline1" and "baseline2" use 3.9 million pairs of Europarl data, in which the difference is that "baseline1" is tuned on the WMT newswire2012 testset, while "baseline2" is tuned on the extracted devset. "DaM1" indicates a combination of Trommons-extracted data and the adaptive data extracted from the Europarl data, and the same for "DaM2"∼"DaM10". The difference in these adaptive data is the amount of data used.

The results of the baselines and adaptive SMT systems are shown in Table 6.

system	BLEU4	TER
baseline0	32.50	53.20
baseline1	30.98	55.83
baseline2	32.57	52.78
DaM1	37.89	48.72
DaM2	37.89	48.60
DaM3	38.56	48.44
DaM4	38.41	48.11
DaM5	39.19	**47.21**
DaM6	38.60	48.32
DaM7	**39.32**	47.45
DaM8	38.91	47.86
DaM9	38.61	48.22
DaM10	38.78	47.92

Table 6: Results of different SMT systems

In Table 6, we can see that:

- "baseline0", "baseline1" and "baseline2" have similar performance, but the data sizes of "baseline1" and "baseline2" are far larger than that of "baseline0", which indicates that there is a significant difference between the Europarl data and the social localisation data in terms of domain.

- The maximum improvement is 6.82 absolute (20.98 relative) BLEU points and is 5.99 absolute (11.26 relative) TER points compared to the "baseline0".

- Immediately, with the increase in training data size selected from the Europarl data using our domain adaptation technique, translation performance correspondingly increases, e.g. the BLEU score is 32.50 for "baseline0", and 37.89 for DaM1 and 39.32 for DaM7, where we see the improvements by 5.39 and 6.82 absolute BLEU points. It is almost the same trend for TER score, where we see the improvements by 4.48 and 5.75 absolute TER points.

- However, when the amount of the data increases to some scale, e.g. from DaM4, translation performance becomes unstable, with performance decreases in some cases, despite the fact that the amount of the data increases.

- Data selection can effectively select appropriate data and keep the training data to a reasonable scale. More importantly, it can be used to quickly build and deploy a relatively high-quality SMT system for practical use.

7 Conclusions

In this paper, we carried out a case study using domain adaptation techniques to augment an SL-SMT system applied in Trommons platform for the English–Spanish language pair. We first examined the file systems in the Trommon platform, and transformed Microsoft Office documents to plain text to be used for SMT building, and then extracted parallel sentences with the alignment rate of 98%. Finally, we used the CED method to select different sizes of data from the Europarl corpus to augment the initial SL-SMT system and achieved improvements of maximum 6.82 absolute BLEU points and 5.99 absolute TER points compared to the baseline.

In future work, we intend to carry out further studies on SL-SMT regarding 1) online domain adaptation: for a more practical SL-SMT system, online domain adaptation can incrementally select ID data to augment SMT system; 2) developing a more effective data selection algorithm to select ID data and localized data; 3) applying our method to new language pairs and domains.

Acknowledgments

We thank the reviewers for their insightful comments. This research is supported by Science Foundation Ireland through the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University, and by Grant 610879 for the Falcon project funded by the European Commission.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain Adaptation via Pseudo InDomain Data Selection. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 20–30, Singapore, Singapore.

Banerjee, P., Naskar, S. K., Roturier, J., and Way, A. (2012). Translation quality-based supplementary data selection by incremental update of translation models. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 149–166, Mumbai, India.

Bicici, E. and Yuret, D. (2011). Instance Selection for Machine Translation using Feature Decay Algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh, UK.

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Jakobsen, A. L. (2011). The process of post-editing: A pilot study. *Copenhagen Studies in Language*, 41:131–142.

Dara, A., van Genabith, J., Liu, Q., Judge, J., and Toral, A. (2014). Active Learning for Post-Editing Based Incrementally Retrained MT. In *Proceedings of the 14th conference of the European Chapter of the ACL (EACL)*, pages 185–189, Gothenburg, Sweden.

Guerberof, A. (2009). Productivity and quality in mt post-editing. In *Proceedings of MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT*, Ottawa, Canada.

Guerberof, A. (2013). What do professional translators think about post-editing? *Journal of Specialised Translation*, 19:75–95.

Haque, R., Penkale, S., and Way, A. (2014). Bilingual Termbank Creation via Log-Likelihood Comparison and Phrase-Based Statistical Machine Translation. In *Proceedings of the 4th International Workshop on Computational Terminology (COLING2014)*, pages 42–51, Dublin, Ireland.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL 2007*, pages 177–180, Prague, Czech Republic.

Lv, Y., Huang, J., and Liu, Q. (2007). Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 343–350, Prague, Czech Republic.

Moore, R. C. and Lewis, W. (2010). Intelligent Selection of Language Model Training Data. In *Proceedings of the Association for Computational Linguistics (ACL) 2010*, pages 220–224, Uppsala, Sweden.

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*, 25(3):197–215.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Schäler, R. (2011). Introducing Social Localisation. `http://www.slideshare.net/TheRosettaFound/social-localisation`.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, USA.

Toral, A. (2013). Hybrid Selection of Language Model Training Data Using Linguistic Information and Perplexity. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 8–12, Sofia, Bulgaria.

Toral, A., Pecina, P., Wang, L., and van Genabith, J. (2014). Linguistically-augmented Perplexity-based Data Selection for Language Models. *Computer Speech and Language*, 32(1):11–26.

Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., and Tron, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP) 2005*, pages 590–596, Borovets, Bulgaria.

Zhechev, V. (2012). Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *Proceedings of AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP2012)*, pages 87–96, San Diego, USA.