Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data?

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier[1], Andy Way[2], Josef van Genabith
CNGL, School of Computing, Dublin City University, Dublin, Ireland
{pbanerjee,snaskar,josef}@computing.dcu.ie
[1] Symantec Limited, Dublin, Ireland
johann_roturier@symantec.com
[2] Applied Language Solutions, Delph, UK
andy.way@appliedlanguage.com

Abstract

This paper reports a set of domain adaptation techniques for improving Statistical Machine Translation (SMT) for user-generated web forum content. We investigate both normalization and supplementary training data acquisition techniques, all guided by the aim of reducing the number of Out-Of-Vocabulary (OOV) items in the target language with respect to the training data. We classify OOVs into a set of types, and address each through dedicated normalization and/or supplementary training material selection-based approaches. We investigate the effect of these methods both in an additive as well as a contrastive scenario. Our findings show that (i) normalization and supplementary training material techniques can be complementary, (ii) for general forum data, fully automatic supplementary training data acquisition can perform as well or sometimes better than semi-automatic normalization (although tackling different types of OOVs) and (iii) for very noisy data, normalization really pays off.

1 Introduction

Web-forums are rich sources of user-generated content on the web. The increasing popularity of technical forums have motivated major IT companies like Symantec to create and support forums around their products and services. For individual users or larger customers, such forums provide an easy source of information and a viable alternative to traditional customer service options. Being a multinational company, Symantec hosts its forums in different languages (English, German, French etc), but currently the content is siloed in each language. Clearly, translating the forums to make information available across languages would be beneficial for Symantec as well as its multilingual customer base. This forms the primary motivation of techniques presented here.

Despite growing interest in translation of forum data (Flournoy and Rueppel, 2010), to date, surprisingly little research has actually focussed on forum data translation (Roturier and Bensadoun, 2011). Compared to professionally edited text, user-generated forum data is often more noisy, taking some liberty with commonly established grammar, punctuation and spelling norms. For our research, we use translation memory (TM) data from Symantec, which is part of their corporate documentation, professionally edited and generally conforming to the Symantec controlled language guidelines. On the other hand, our target data (forum) is only lightly moderated and does not conform to any publication quality guidelines. Hence despite being from the same IT domain, there is a significant difference in style between the training and the test data. In this paper, we focus our efforts on systematically reducing this difference through the use of both normalization and supplementary training material acquisition techniques.

Our research was conducted on English to German (En–De) and English to French (En–Fr) language directions. To identify the differences between the TM and forum data, we focus on the OOV words in the English forum data with respect to the source side (English) of the TM data. We classify OOVs into different categories which require independent attention. In order to optimally handle each individual category, different tech-

niques were developed to make the forum-based test sets better resemble the training data. For the *first category – containing tokens such as URLs, paths, registry entries, and memory addresses* – regular expressions were used to capture the tokens and replace them with unique place-holders. The *second category included valid words inadvertently fused by punctuation characters (especially '.')* which required a training data-guided splitting technique. The *third category comprising spelling errors* were handled by an off-the-shelf automatic spell checker. Additionally the spell checker was trained with 'in domain' data to make it aware of the domain-specific terms to improve the quality of spell checking. For the *fourth category of OOVs – valid words not occurring in the training data* – various supplementary 'out-of-domain' bitext training data were automatically searched. For every OOV in this category, parallel sentence pairs from different 'out-of-domain' data were added to the 'in-domain' training data to improve the coverage of the translation models.

While improving translation quality by reducing OOVs is the primary objective of our research, we are particularly interested in the effect of spell checking on translation quality of forum data with various degrees of noise. Furthermore, we compare the relative improvements provided by the normalization to supplementary data selection to justify the effectiveness of the respective techniques. The rest of the paper is organized as follows: Section 2 briefly reviews relevant related work. Section 3 provides a detailed discussion on the normalization techniques as well as the acquisition of supplementary training material. Section 4 presents the datasets and the experiments and corresponding results, followed by our conclusions and pointers to future work in Section 5.

2 Related Work

The technique of using 'out-of-domain' datasets to supplement 'in-domain' training data has been widely used in domain adaptation of SMT. Information retrieval techniques were used by Eck et al. (2004) to propose a language model adaptation technique for SMT. Hildebrand et al. (2005) utilized this approach to select similar sentences from available bitext to adapt translation models, which improved translation performance. Habash (2008) used spelling expansion, morphological expansion, dictionary term expansion and proper name transliteration to enhance or reuse existing phrase table entries to handle OOVs in Arabic–English MT. More recently an effort to adapt MT by mining bilingual dictionaries from comparable corpora using untranslated OOV words was carried out by Daume III and Jagarlamudi (2011).

Our current line of work is related to the work reported in Daume III and Jagarlamudi (2011) and that of Habash (2008). In our case, however, the target domain (web-forum) is different from the training data (Symantec TMs) more in terms of style rather than actual domain (Banerjee et al., 2011). Secondly, in contrast to mining comparable data for bilingual dictionary extraction (Daume III and Jagarlamudi, 2011), we exploit sentence pairs from available parallel training data to handle untranslated OOVs. Moreover, mining supplementary parallel data guided by OOVs is used as a technique complementing the normalization-based approaches to reduce specific types of OOVs in the target domain. We classify OOVs into different categories and treat each of them separately. In contrast to extending the phrase table entries (Habash, 2008) our normalization methods mostly comprise pre- and post-processing techniques. Finally we also present a comparison between the normalization and supplementary training data acquisition techniques for different error density-based scenarios of the target domain. To the best of our knowledge, the use of 'domain-adapted' spell checkers to reduce OOV rates in the target domain is novel, and is one of the other main contributions of the paper.

3 Normalization and Supplementary Data Selection Techniques

This section introduces the datasets used for the experiments followed by the adaptation techniques used in the experiments.

3.1 Datasets

The primary training data for our experiments consisted of En–De and En–Fr bilingual datasets in the form of Symantec TMs. Monolingual Symantec forum posts in German and French along with the target side of the TM training data served as language modelling data. In addition, we also had a collection of posts from the original Symantec English forums acquired over a period of two years which formed the basis of our OOV category estimation. The development (dev) and test sets used

in our experiments were randomly selected from this particular data set. Table 1 reports the amount of data used for all our experiments.

	Data Set	En–De	En–Fr
Bi-text	Symantec TM	832,723	702,267
	Development Set	500	500
	Test 1	2,022	2,022
	Test 2	600	600
Monolingual	English Forum		1,129,749
	German Forum		42,521
	French Forum		41,283

Table 1: Number of Sentences for training, development and test sets, and forum data sets

As reported in Table 1, we used two different test sets, for our experiments. The first one (Test-1) was randomly chosen from the English forum data. Since one of our objectives was also to investigate a scenario with a high density of spelling errors, typical for some forum posts, the second test set (Test-2) was selected to simulate a higher proportion of noise (approximately one spelling error in every two test set sentences). This was achieved by flagging the remaining forum dataset (after removing the Test-1 sentences), using an automatic spell checker, and randomly selecting sentences with spelling errors followed by a manual review. Both these test sets were manually translated following basic guidelines for quality assurance. The randomly chosen dev set was translated using Google Translate,[1] and manually post-edited by professional translators following guidelines[2] for achieving 'good enough quality'.

3.2 OOV Categorization

The remaining (after dev and test set selection) English forum data, comprising over 1.13M sentences (around 17.5M words), were used to compute OOV words in the forum domain with respect to the training data, using a unigram language model estimated on the source side of the training data. Manual inspection of the OOV word list identifies the following general categories:

1. Maskable Tokens (MASK): URLs, paths, registry entries, email addresses, memory locations, date and time tokens and IP addresses or version numbers.
2. Fused Words (FW): Two or more valid tokens concatenated using punctuation marks like '.'.
3. Spelling Errors (SPERR): Spelling errors or typos.
4. Valid Words (VAL): Valid words not occurring in the training data.
5. Non-Translatable (NTR): Tokens comprising standalone product and service names and numbers (not part of Category-1 tokens) which ideally should not be translated.

Table 2 depicts the percentage of the OOV word categories in the English forum data and the two test sets with respect to the En–De and En–Fr TM-based source data sets. Comparing the category-wise percentage figures on the two test sets (Test-1 and Test-2) clearly show the distribution of the categories in Test-1 is similar to that of the original Forums. Test-2 shows a higher percentage of SPERR tokens as it had been consciously designed to have high spelling error density. The figures also depict the relative importance of the specific OOV categories in forum-style data, with non-translatable (NTR) and maskable tokens (MASK) covering nearly 75% of the OOV range.

OOV	En–De			En–Fr		
Type	Forum	Test-1	Test-2	Forum	Test-1	Test-2
MASK	25.68	21.33	9.93	25.47	19.43	9.83
FW	8.89	4.11	2.05	8.75	3.71	2.00
SPERR	10.41	12.64	52.91	10.45	12.29	52.67
VAL	6.38	14.06	12.33	6.74	18.86	12.17
NTR	48.64	47.87	22.77	48.60	45.71	23.33

Table 2: Category-based percentage of OOVs in the English forum and two test data sets

Different normalization techniques used to independently address each of these OOV categories are detailed below.

3.3 Regular Expression-based Normalization

For the normalization of MASK OOVs we developed a set of regular expressions to identify tokens. These were replaced with unique placeholders. These replacements were then applied uniformly over all data sets (TM and forum) in a pre-processing step. Most of the tokens in this category were multi-word tokens, and this method allowed them to be treated as single tokens during the translation process. This not only helped in maintaining the internal ordering of words within such tokens but also ensured that none of the terms within such a token were translated.

3.4 Fused Word Splitting

To handle FW tokens which comprise two or more valid words fused using a period ('.') symbol, we

[1] http://translate.google.com/

[2] http://www.translationautomation.com/machine-translation-post-editing-guidelines.html

identified all tokens which had a period symbol flanked by alphabetic characters. However, since a large number of valid file names, website names or abbreviations (e.g. N.I.S., explorer.exe, shopping.aol.com, etc.) were also identified, we used heuristics based on the training data to identify the valid ones. Lists of known file extensions (e.g. exe, jar, pdf, etc.) and website domain extensions (e.g. com, edu, net, gov, co.uk, etc.) were used to filter out file names and website names. Finally we used a dictionary built on the training data. Every split was validated against the dictionary, with the constraint that all its constituent splits had to occur in this dictionary. This normalization was only applied on the dev and test sets as the TM training data was assumed to be clean of such fused words.

3.5 spell checker-based Normalization

A considerable amount of the OOVs in the unnormalized forum data comprise spelling errors or typos (SPERR). We used an off-the-shelf spell checker (cf. Section 4.2) to identify and correct these tokens so that they mapped to valid words (preferably in the training data). While the ready-to-use spell checker worked well for most of the spelling errors in general-purpose English words, it flagged a lot of 'in-domain' (technical) words. Hence we adapted the spell checker to the domain. This was achieved by generating glossary lists from the source side of the TMs and adding them to the spell checker dictionary. Furthermore, the spell checking models had to be retrained using the source side of 'in-domain' data from TMs. The adaptation of the spell checker helped us to eliminate most of the false positives flagged by the original unadapted spell checker. The errors flagged by the spell checker were replaced with the highest ranking suggestion from the spell checker. As in Section 3.4, the spelling corrections were applied only to the test sets to ensure a reduction in the number of spelling error-based OOVs.

3.6 Supplementary Data Selection

To take care of the VAL tokens which are valid words but absent in the training data, we explored techniques of mining supplementary data to improve the chances of successfully translating these tokens. We used the following freely available parallel data collections as potential sources of supplementary data:

1. Europarl (Koehn, 2005): Parallel corpus comprising of the proceedings of the European Parliament.
2. News Commentary Corpus: Released as a part of the WMT 2011 Translation Task.[3]
3. OpenOffice Corpus: Parallel documentation of the Office package from OpenOffice.org, released as part of the OPUS corpus (Tiedemann, 2009).
4. KDE4 Corpus: A parallel corpus of the KDE4 localization files released as part of OPUS.
5. PHP Corpus: Parallel corpus generated from multilingual PHP manuals also released as part of OPUS.
6. OpenSubtitles2011 Corpus:[4] A collection of documents released as part of OPUS.
7. EMEA Corpus: A parallel corpus from the European Medical Agency also released as part of OPUS corpus.

To select relevant parallel data, we queried each of the parallel corpora with the VAL OOV words and added sentence pairs containing the OOVs into the existing 'in-domain' parallel corpora. During the selection process, the number of parallel sentences selected for any particular OOV item was restricted to a threshold of 500 for En–De and 67 for En–Fr. This was done to limit the size of the selected 'out-of-domain' supplementary data such that it did not exceed the size of the TM-based (in-domain) training data. The target sentences of the selected parallel data were added to the language model to ensure language model adaptation. This process allowed us to cover 87.55% and 92.13% of VAL OOVs for En–De and En–Fr language pairs, respectively.

3.7 OOV Tokens Unsuitable for Translation

The last remaining category of OOVs (NTR) represents tokens for which translation was usually unnecessary. Most of these comprised product or service names, names of the forum users or numeric tokens. This class of tokens was not explicitly handled under the assumption that due to their absence from the training data (and hence from the phrase table), they would be preserved during the translation process in the standard SMT setup.

[3] http://www.statmt.org/wmt11/translation-task.html
[4] http://www.opensubtitles.org/

4 Experiments and Results

4.1 Pre- and Post-Processing

Prior to training, all the bilingual and monolingual data were subjected to tokenization and lower casing using the standard Moses pre-processing scripts. However, for the regular expression-based normalization, the standard tokenizer is slightly modified to ensure that unique placeholders (Section 3.3) are not tokenized. During the replacement process a mapping is maintained between the unique placeholders, the line number and the actual token replaced. This mapping file is used later in the post-processing step to substitute the actual tokens in the position of the unique placeholders. For target sentences having multiple placeholders of the same type, the corresponding actual tokens are replaced in the order in which they appeared in the source.

4.2 Tools

For all our translation experiments we used OpenMaTrEx (Dandapat et al., 2010), an open source SMT system which wraps the standard log-linear phrase-based SMT system Moses (Koehn et al., 2007). Word alignment was performed with Giza++ (Och and Ney, 2003). The phrase and reordering tables were built on the word alignments using the Moses training script. The feature weights for the log-linear combination of the feature functions were tuned using Minimum Error Rate Training (Och, 2003) on the devset in terms of BLEU (Papineni et al., 2002). We used 5-gram language models in all our experiments created using the IRSTLM (Federico et al., 2008) language modelling toolkit using Modified Kneser-Ney smoothing. Results of translations in every phase of our experiments were evaluated using BLEU and TER (Snover et al., 2006).

For the spell checking task we used a combination of two off-the-shelf spelling correction toolkits. Using the 'After the Deadline toolkit' (AtD)[5] as our primary spell checker, we also used a Java wrapper on Google's spellchecking API[6] to supplement the AtD spell checking results. However, the 'in-domain' adaptation of the spell checker (Section 3.5) could only be achieved for the AtD spell checker.

[5]http://open.afterthedeadline.com/

[6]http://www.google.com/tbproxy/spell?lang=en&hl=en

4.3 Experimental Results

Table 3 shows the different BLEU and TER scores for translations subject to each category of normalization and supplementary data selection, along with the percentage of OOV word reduction they result in, for both the test sets under consideration. The last row of the table reports the results for translating only regular expression-based normalized test sets (without the other normalizations) using supplementary training data enhanced models.

The experiments were carried out in five different phases, each focussing on reducing one category of OOV words in the English forum data. For the baseline translation and language models, the TM and forum data was subjected to only basic clean-up such as dropping empty lines and very long sentences (more than 100 tokens). The baseline testsets were then subjected to the following adaptations in a cumulative step-by-step manner:

1. Regex: Regular Expression-based normalization for the reduction of MASK OOVs.
2. Wrd-Split: Heuristic-based tokenization for normalization of FW OOVs.
3. Spell-Chk: Off-the-shelf spell checking based normalization for reducing SPERR.
4. Adapted-Spell-Chk (Ada SpChk): spell checking using domain adapted spell checkers to reduce false positive flags.
5. Sup-data: Supplementary data selection and addition to enrich existing models to reduce VAL OOVs.

The final experimental step (Regex+Sup) did not involve any specific normalization, but was rather performed to investigate the effect of supplementary data selection on regex-based normalized test sets without any other normalizations.

As the results in Table 3 show, regular expression-based normalization results in a 0.55 absolute (2.12% relative) BLEU point improvement in En–De translations and a 0.66 absolute (1.93% relative) BLEU point improvement for En–Fr translations for Test-1. For Test-2, the improvements are 0.31 absolute (1.45% relative) BLEU points and 0.38 absolute (1.26% relative) BLEU points for En–De and En–Fr, respectively. While the Test-1 improvements are statistically significant at p=0.05 level using bootstrap resampling (Koehn, 2004), the Test-2 improvements are not statistically significant. The TER scores also

Normaliz-ation	En–De						En–Fr					
	Test-1			Test-2			Test-1			Test-2		
	OOV	BLEU	TER	OOV	BLEU	TER	OOV	BLEU	TER	OOV	BLEU	TER
Baseline	–	25.98	0.6407	–	21.32	0.6361	–	34.14	0.5250	–	30.27	0.5405
Regex	21.33	26.53*	0.6372	9.42	21.63	0.6332	19.43	34.80*	0.5179	9.67	30.65	0.5402
Wrd-Split	3.48	26.59	0.6380	1.54	21.68*	0.6284	3.14	34.89	0.5178	1.50	30.77*	0.5386
Spell-Chk	8.06	26.78	0.6365	37.16	22.50*	0.6279	8.57	35.10	0.5158	36.17	31.60*	0.5303
Ada-SpChk	4.27	26.92	0.6299	11.30	23.17*	0.6174	3.57	35.33	0.5121	11.00	32.28*	0.5128
Sup-data	13.74	27.86*	0.6207	13.53	24.08*	0.5923	17.43	36.04*	0.5024	15.17	33.75*	0.5043
Regex-Sup	13.74	27.45	0.6242	13.53	23.01	0.6191	17.43	35.55	0.5068	15.17	31.96	0.5178

Table 3: Translation Results after normalization and supplementary data selection. The OOV column indicate the percentage of total OOVs reduced in each step. * denote statistically significant improvement over the scores in previous row.

show a decreasing trend which also suggest translation quality improvement. The reason behind this may be attributed to the larger percentage of category-1 tokens in Test-1 compared to Test-2. The number of OOV words is reduced by 135 and 136 on Test-1 and 55 and 58 on Test-2 with respect to different training data sets. The improvements result from the fact that this normalization helps to maintain intra word ordering within MASK tokens and avoid translation of constituent sub-tokens. The first example in Table 4 clearly depicts this particular behaviour for MASK tokens.

Using the fused word splitting technique on the regex-processed testsets, the scores improve only by 0.06 absolute (0.23% relative) BLEU points and 0.09 (0.26% relative) absolute BLEU points on Test-1 over the previous normalization scores, for En–Fr and En–De respectively. For Test-2 the improvements are 0.05 absolute (0.23% relative) BLEU points and 0.12 absolute (0.39%) BLEU points for En–De and En–Fr translations, respectively. Despite the marginal improvement, the improvements for Test-2 were statistically significant at p=0.05 level. Improvements in Test-1 were not significant. The reason for the marginal improvement becomes apparent when observing the low percentage of OOV's (Table 3) reduced by this mechanism. However, the percentage of category-2 tokens in test-2 is nearly double that of Test-1 which may explain the statistical significance of the improvements gained.

As expected, handling the spelling errors using spell checkers had a profound effect on the reduction of OOV words for the high density spelling error testset, Test-2. Using the adapted spell checker on this test set, we achieve an improvement of 1.49 absolute (6.87% relative) BLEU points for En–De and 1.51 absolute (4.9%) BLEU points for En–Fr translations. This corresponds to a total reduction (combining reductions for unadapted and adapted spell checking) of 283 OOVs for both En–De and En–Fr test sets. The overall improvement when using spell checkers over the previous normalization results were statistically significant at the p=0.05 level. However, for Test-1, with spelling error density reflecting that of average forum data, the improvements are much lower. Adapted spell checking results in a total improvement of 0.33 absolute (1.24% relative) BLEU points for En–De and 0.44 absolute (1,26% relative) BLEU points for En–Fr translations. These are not statistically significant and correspond to a reduction of 78 and 85 OOVs for En–De and En–Fr test sets, respectively. The TER scores also reflect the same level of improvements across the two different test sets.

The fourth phase of experiments, where different parallel data sources are mined guided by the list of VAL OOV words, results in further reduction in the OOV rates and improvement in translation scores. The guided selection process improves the scores by 0.94 absolute (3.49% relative) and 0.71 absolute (2.01% relative) BLEU points for En–De and En–Fr translations, respectively on Test-1. For Test-2 the improvement figures are 0.91 absolute (3.93% relative) BLEU points and 1.47 absolute (4.55% relative) BLEU points for En–De and En–Fr translation, respectively, over the previous normalization results. The TER scores also show similar improvements for both language pairs and test sets. All improvements are statistically significant at the p=0.05 level. Furthermore, this technique further reduces the number of OOVs by 79 for the En–De test set and 91 counts for the En–Fr on Test-2. The corresponding reductions for Test-1 are 87 and 122 for En–De and En–Fr, respectively.

In summary, using supplementary data selection techniques to complement the normalization resulted in statistically significant overall improvements of 1.88 absolute (7.24% relative) and 1.9 absolute (5.57% relative) BLEU points over the baseline scores on Test-1. On Test-2, the im-

provements were 2.76 absolute (12.95% relative) and 3.48 absolute (11.49% relative) BLEU points for En–De and En–Fr translations, respectively. Translating the regex-normalized test sets (without word splitting and spell checking) with the supplementary data-enhanced models, we aimed to assess the impact of supplementary data selection technique in contrast to that of the normalization methods. For Test-1, the results show that this process results in scores slightly better (0.53 absolute BLEU on En–De and 0.22 absolute BLEU for En–Fr) than those achieved by complete normalization (adapted spell checking scores, row 5 in Table 4.3). For Test-2 however, the scores are lower than the adapted spell checking scores by 0.16 and 0.32 absolute BLEU points for En–De and En–Fr, respectively. Overall results clearly show that for general forum data (with average spelling error density), fully automatic supplementary training data acquisition can perform as well and sometimes better than semi-automatic normalization although they target different types of OOVs. Finally for very noisy data, normalization complemented with supplementary data selection really pays off.

Type	Sentence
Src	5 . click on the folder button and navigate to **c : \documents and settings \all users \application data **and select the carbonite folder
Ref	5. klicken sie auf die ordnerschaltfläche und öffnen sie den ordner " **c : \documents and settings \all users \application data \carbonite "**
Baseline	5. klicken sie auf den ordner " und navigieren sie zu **c : \dokumente und einstellungen \alle benutzer \anwendungsdaten ** und wählen sie die carbonite ordner
Regex	5. klicken sie auf die schaltfläche " und wechseln sie zum ordner **c : \documents and settings \all users \application data \carbonite** und wählen sie die carbonite ordners
Src	re : nis09 did not detect 8 threats & 23 infected **objects.and** 16 suspicious objects ?
Ref	re : nis09 n' a pas détecté 8 menaces , 23 **objets infectés et** 16 objets suspects ?
Baseline	re : nis09 n' a pas détecter 8 menaces et 23 infecté **objects.and** 16 les objets ?
Wrd-Split	re : nis09 n' a pas détecter 8 menaces et 23 **infecté objets . et** 16 les objets ?
Src	and no for **somthing completly** different .
Ref	und nun zu **etwas völlig** anderem .
Baseline	und keine für **somthing completly** anders .
Spck	und nicht für **etwas völlig** anders .
Src	pretty **disappointed** with nis parental control not blocking websites on blocked list as well as through their category of websites to block .
Ref	je suis assez **déçu** que le contrôle parental de nis ne bloque pas les sites web figurant dans la liste bloqués aussi bien que ceux de la catégorie des sites web à bloquer .
Baseline	assez **disappointed** avec contrôle parental de nis pas le blocage de sites web sur liste bloqués ainsi que par l' intermédiaire de leur catégorie de sites web à bloquer .
Sup	assez **déçu** de contrôle parental de nis pas le blocage de sites web sur liste bloqués ainsi que dans leur catégorie de sites web à bloquer .

Table 4: Translation examples for each normalization and supplementary data selection Technique

In order to substantiate the improvements observed on the automatic evaluation scores, we present some examples from our test sets (both Test-1 and 2), to depict how the normalization or data selection methods actually affect the translations. Table 4 presents 4 different examples of translations each highlighting the effect of a single normalization or data selection technique. The first example clearly shows how regular expression-based masking allows internal parts of the path structure to be left untranslated, unlike in the baseline set-up. The second sentence (row 5) is an example of the fused word splitting technique enabling better translation of the token 'objects.and' which had been treated as an OOV in the baseline. The third example (rows 9-12) highlights the effect of spell checking on the translation quality of the source sentence. Automatic spell checking changes the tokens 'somthing completly' into 'something completely' thereby allowing them to be translated. The final set of sentences is an example of how supplementary data selection allows the translation of the valid yet OOV word 'disappointed' appearing in the source sentence. As is evident from the examples, the normalization techniques discussed in the paper do work towards better translations for sentences with specific OOV types. However, the relative densities of each type leads to varied improvements in scores reported in Table 4.3.

5 Conclusion and Future Work

In this paper we have explored a set of normalization techniques to achieve better translation quality for user-generated forum content. We have shown that supplementary data selection techniques positively complement normalization in terms of translation quality. For test data with spelling error density representative of the overall forum data (Test-1), supplementary data selection on its own can produce improvements similar to those achieved through normalisation (targeting different OOVs). While data normalization carried out at the level reported in this paper (with different OOV categories and different normalisation approaches for each) is a semi-automatic process which requires some manual analysis, supplementary data selection is fully automatic and involves much less overall effort. Thus, for moderately noisy datasets (such as Test-1), normalization may not always be worth the effort. For more noisy datasets (e.g. Test-2) however, normalization does improve translation quality more effectively than data supplementation.

In this research, the classification of OOV words was done in a semi-automatic fashion. Using automatic classification techniques to identify the different categories in OOV words would be one of the prime future directions here. Furthermore, a detailed investigation of the individual contributions of multiple resources used for supplementary

data selection is required to better understand the cause of the improvements in scores. Finally we would also like to work towards developing automatic threshold detection techniques for optimal supplementary data selection.

Acknowledgments

This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. We thank the reviewers for their insightful comments.

References

Banerjee, Pratyush, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2011. Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component Level Mixture Modelling. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 285–292, Xiamen, China.

Dandapat, S., M. L. Forcada, D. Groves, S. Penkale, J. Tinsley, and A. Way. 2010. OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based Machine Translation System. In *Proceedings of the 7th International Conference on Natural Language Processing (IceTAL 2010)*, page 121–126, Reykjavík, Iceland.

Daume III, Hal and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, Portland, Oregon, USA.

Eck, Matthias, Stephan Vogel, and Alex Waibel. 2004. Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of 4th International Conference on Language Resources and Evaluation, (LREC 2004)*, pages 327–330, Lisbon, Portugal.

Federico, Marcello, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech 2008: 9th Annual Conference of the International Speech Communication Association*, pages 1618–1621, Brisbane, Australia.

Flournoy, Raymond and Jeff Rueppel. 2010. One Technology : Many Solutions. In *AMTA 2010: Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, pages 6–12, Denver, Colorado, USA.

Habash, Nizar. 2008. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 57–60, Columbus, Ohio.

Hildebrand, Almut Silja, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *10th EAMT Conference: Practical Applications of Machine Translation, Conference Proceedings*, pages 119–125, Budapest, Hungary.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, Phillipe. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP 2004)*, pages 388–395, Barcelona, Spain.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit X: The 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167, Sapporo, Japan.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics, (ACL 2002)*, pages 311–318, Philadelphia, Pennsylvania.

Roturier, Johann and Anthony Bensadoun. 2011. Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 244–251, Xiamen,China.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA.

Tiedemann, Jörg. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. pages 237–248.