A Framework for Translating SMS Messages

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Ron Shacham
AT&T Labs
1 AT&T Way, Bedminster, NJ 07921
vkumar,jchen,srini,rshacham@research.att.com

Abstract

Short Messaging Service (SMS) has become a popular form of communication. While it is predominantly used for monolingual communication, it can be extremely useful for facilitating cross-lingual communication through statistical machine translation. In this work we present an application of statistical machine translation to SMS messages. We decouple the SMS translation task into normalization followed by translation so that one can exploit existing bitext resources and present a novel unsupervised normalization approach using distributed representation of words learned through neural networks. We describe several surrogate data that are good approximations to real SMS data feeds and use a hybrid translation approach using finite-state transducers. Both objective and subjective evaluation indicate that our approach is highly suitable for translating SMS messages.

1 Introduction

The preferred form of communication has been changing over time with advances in communication technology. The majority of the world's population now owns a mobile device and an ever increasing fraction of users are resorting to Short Message Service (SMS) as the primary form of communication.

SMS offers an easy, convenient and condensed form of communication that is being embraced by the younger demographic. Due to the inherent limit in the length of a message that can be transmitted, SMS users have adopted several shorthand notations to compress the message; some that have become standardized and many that are invented constantly. While SMS is predominantly used in a monolingual mode, it has the potential to connect people speaking different languages. However, translating SMS messages has several challenges ranging from the procurement of data in this domain to dealing with noisy text (abbreviations, spelling errors, lack of punctuation, etc.) that is typically detrimental to translation quality. In this work we address all the elements involved in building a cross-lingual SMS service that spans data acquisition, normalization, translation modeling, messaging infrastructure and user trial.

The rest of the paper is organized as follows. In Section 4, we present a variety of channels through which we compiled SMS data followed by a description of our pipeline in Section 5 that includes normalization, phrase segmentation and machine translation. Finally, we describe a SMS translation service built using our pipeline in Section 6 along with results from a user trial. We provide some discussion in Section 7 and conclude in Section 8.

2 Related Work

One of the main challenges of building a machine translation system for SMS messages is the lack of training data in this domain. Typically, there are several legal restrictions in using consumer SMS data that precludes one from either using it completely or forces one to use it in limited capacity. Only a handful of such corpora are publicly available on the Web (Chen and Kan, 2013; Fairon and Paumier, 2006; Treurniet et al., 2012; Sanders, 2012; Tagg, 2009); they are limited in size and restricted to a few language pairs.

The NUS SMS corpus (Chen and Kan, 2013) is probably the largest English SMS corpus consisting of around 41000 messages. However, these messages are characteristic of Singaporean chat lingo and not an accurate reflection of SMS style in other parts of the world. A corpus of 30000 French SMS messages

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 974–983, Dublin, Ireland, August 23-29 2014.

was collected in (Fairon and Paumier, 2006) to study the idiosyncrasies of SMS language in comparison with standard French. More recently, (Pennell and Liu, 2011) have used twitter data as a surrogate for SMS messages. Most of these previous efforts have focused on normalization, i.e., translation of SMS text to canonical text while we are interested in translating SMS messages from one language into another (Eidelman et al., 2011).

Several works have addressed the problem of normalizing SMS text. A majority of these works have used statistical machine translation (character-level) to translate SMS text into standard text (Pennell and Liu, 2011; Aw et al., 2009; Kobus et al., 2008). (Beaufort et al., 2010) used a finite-state framework to learn the mapping between SMS and canonical form. A beam search decoder for normalizing social media text was presented in (Wang and Tou Ng, 2013). All these approaches rely on supervised training data to train the normalization model. In contrast, we use an unsupervised approach to learn the normalization lexicon of word forms in SMS to standard text.

While several works have addressed the problem of normalizing SMS using machine translation, there has been little to no work on the translation of SMS messages across languages on a large scale. Machine translation of instant messages from English-to-Spanish was proposed in (Bangalore et al., 2002) where multiple translation hypotheses from several off-the-shelf translation engines were combined using consensus decoding. However, the approach did not consider any specific strategies for normalization and the fidelity of training bitext is questionable since it was obtained using automatic machine translation. Several products that enable multilingual communication with the aid of machine translation in conventional chat, email, etc., are available in the market. However, most of these models are trained on relatively clean bitext.

3 Problem Formulation

The objective in SMS translation is to translate a foreign sentence $\mathbf{f}^{sms} = f_1^{sms}, \cdots, f_J^{sms}$ into target (English) sentence $\mathbf{e} = e_1^I = e_1, \cdots, e_I$. In general it is hard to procure such SMS bitext due to lack of data and high cost of annotation. However, we typically have access to bitext in non-SMS domain. Let $\mathbf{f} = f_1, \cdots, f_J$ be the normalized version of the SMS input sentence. Given $\mathbf{f}^{sms}$, we choose the sentence with highest probability among all possible target sentences,

$$\hat{\mathbf{e}}(\mathbf{f}^{sms}) = \arg\max_{\mathbf{e}} \{\mathrm{P}(\mathbf{e}|\mathbf{f}^{sms})\} \tag{1}$$

$$P(\mathbf{e}|\mathbf{f}^{sms}) \approx P(\mathbf{e}) \sum_{\mathbf{f}} P(\mathbf{f}^{sms}, \mathbf{f}|\mathbf{e}) \tag{2}$$

$$= P(\mathbf{e}) \sum_{\mathbf{f}} P(\mathbf{f}^{sms}|\mathbf{f}, \mathbf{e}) P(\mathbf{f}|\mathbf{e}) \tag{3}$$

If one applies the max-sum approximation and assumes that $P(\mathbf{f}^{sms}|\mathbf{f}, \mathbf{e})$ is independent of $\mathbf{e}$,

$$\hat{\mathbf{e}}(\mathbf{f}^{sms}) = \arg\max_{\mathbf{e}} P(\mathbf{f}^*|\mathbf{e}) P(\mathbf{e}) \tag{4}$$

where $\mathbf{f}^* = \arg\max_{\mathbf{f}} P(\mathbf{f}^{sms}|\mathbf{f})$. Hence, the SMS translation problem can be decoupled into normalization followed by statistical machine translation[1].

4 Data

Typically, one has access to a large corpus of general bitext $\{\mathbf{f}, \mathbf{e}\}$ while data from the SMS domain $\{\mathbf{f}^{sms}, \mathbf{e}\}$ is sparse. Compiling a large corpus of SMS messages is not straightforward as there are several restrictions on the use of consumer SMS data. We are not aware of any large monolingual or bilingual corpus of true SMS messages besides those mentioned in Section 2. To compile a corpus of SMS messages, we used three sources of data: transcriptions of speech-based SMS collected through

[1] One can also use a lattice output from the normalization to jointly optimize over $\mathbf{e}$ and $\mathbf{f}$

smartphones, data collected through Amazon Mechanical Turk[2] and Twitter[3] as a surrogate for SMS-like messages. We describe the composition of each of these data sources in the following subsections.

Corpus	Message	#count	Corpus	Message	#count
Speech SMS	*i love you*	988157	Amazon Mechanical Turk	*ily2*	N/A
	hello	881635		*n a meeting*	
	hi	607536		*check facebook*	
	how are you	470999		*kewl*	
	what's up	251044		*call u n a few*	
	what are you doing	218289	Twitter	*lol*	472556
	where are you	191912		*haha*	232428
	call	191430		*lmao*	102018
	lol	105618		*omg*	709504
	how's it going	102977		*thanks for the rt*	300254

Table 1: Examples of English messages collected from various sources in this work

4.1 Speech-based SMS

In the absence of access to a real feed of SMS messages, we used transcription of speech-based SMS messages collected through a smartphone application. A majority of these messages were collected while the users used the application in their cars. We had access to a total of 41.3 million English and 2.4 million Spanish automatic transcriptions. To avoid the use of erroneous transcripts, we sorted the messages by frequency and manually translated the top 40,000 English and 10,000 Spanish messages, respectively. Our final English-Spanish bitext corpus from this source of data consisted of 50,000 parallel sentences. Table 1 shows the high frequency messages in this dataset.

4.2 Amazon Mechanical Turk

The SMS messages from speech-based interaction does not consist of any shorthands or orthographic errors as the decoding vocabulary of the automatic speech recognizer is fixed. We posted a task on Amazon Mechanical Turk, where we took the speech-based SMS messages and asked the turkers to enter three responses to each message as they would on a smartphone. We iteratively posted the responses from the turkers as messages to obtain more messages. We obtained a total of 1000 messages in English and Spanish, respectively. Unlike the speech data, the responses contained several shorthands.

4.3 Twitter

Twitter is used by a large number of users for broadcasting messages, opinions, etc. The language used in Twitter is similar to SMS and contains plenty of shorthands, spelling errors even though it is typically not directed towards another individual. We compiled a data set of Twitter messages that we subsequently translated to obtain a bilingual corpus. We used the Twitter4j API[4] to stream Twitter data for a set of keywords (function words) over a week. The raw data consisted of roughly 106 million tweets. Subsequently, we performed some basic normalization (removal of @user, #tags, filtering advertisements, web addresses) to obtain SMS-like tweets. Finally, we sorted the data by frequency and picked the top 10000 tweets. Eliminating the tweets present in either of the two previous sources resulted in 6790 messages that we manually translated.

5 Framework

The user input is first stripped of any accents (Spanish), segmented into short chunks using an automatic punctuation classifier. Subsequently, any shorthand in the message is expanded out using expansion dictionaries (constructed manually and automatically) and finally translated using a phrase-based translation

[2]https://www.mturk.com
[3]https://twitter.com
[4]http://twitter4j.org/en/

model. Our framework allows the use of confusion networks in case of ambiguous shorthand expansions. We describe each component of the pipeline in detail in the following sections.

5.1 Tokenization

Our initial analysis of SMS messages from users, especially in Spanish indicated that while some users use accented characters in orthography, several others omit it for the sake of faster responses and convenience. Hence, we decided to train all our models on unaccented characters. Given a message, we convert all accented characters to their corresponding unaccented forms, e.g., baño → bano, followed by lowercasing of all characters. We do not perform any other kind of tokenization.

5.2 Unsupervised SMS Normalization

In Section 5.2, we described a static lookup table for expanding abbreviations and shorthands typically encountered in SMS messages, e.g., *4ever→forever*. While a static lookup table provides a reasonable way of handling common SMS abbreviations, it has limited coverage. In order to build a larger normalization lexicon, we used distributed representation of words to induce the lexicon in an unsupervised manner. Distributed word representations (Bengio et al., 2003; Collobert and Weston, 2008; Turian et al., 2010) induced through deep neural networks have been shown to be useful in several natural language processing applications. We use the notion of distributional similarity that is automatically induced through the word representations for learning automatic normalization lexicons.

Canonical form	Noisy form
love	loveeee, loveeeee, looove, love, wuv, wove, love, laffff, love, wuvvv, luhhhh, love, luvvv, luv
starbucks	starbs, sbucks
once	oncee, 1ce
tomorrow	tmrw, tomorrow, 2moro, tmrrw, tomarrow, tomoro, tomoz, 2mrw, tmr, tm, tmwr, 2mm, tmw, 2morro
forever	foreva, 5ever, foreverrrr, forver, foreeverrr, 4ever, 5eva, 4eva, foreevaa, forevs, foreve
because	cause, cos, coz, 'cos, 'cause, bc, because, becuz, bcuz, cuz, bcus, bcoz, because
homework	hwk, hw, hmwk, hmwrk, hmw, homeworkk, homwork, hmk, honework, homeowork
igualmente	igualmentee, igualment, iwalmente
siempre	simpre, siempre, 100pre, siempre, ciempre, siempre, siiempre, siemore, siempr, siemre, siempe
adios	adi, a10, adio
contigo	contigoo, cntigo, conmigo, contigoooo, kontigo, conmigoo, conmiqo
demasiado	demaciado, demasido, demasiademente, demasiao

Table 2: Examples from the unsupervised normalization lexicon induced through deep learning

We started with the 106 million tweets described in Section 4.3 and used a deep neural network identical to that used in (Collobert and Weston, 2008), i.e., the network consisted of a lookup table, hidden layer with 100 nodes and a linear layer with one output. However, we used a context of 5 words and corrupted the centre word instead of the last word to learn the distributed representations. We performed stochastic gradient minimization over 1000 epochs on the twitter data. Subsequently, we took the English and Spanish vocabularies in our translation model and found the 50 nearest neighbors using cosine distance for each word. We trained the above representations using the Torch toolkit (Collobert et al., 2011).

Feature	English		Spanish	
dimension	Precision	Recall	Precision	Recall
100	70.4	97.4	69.8	97.3
200	72.2	97.5	79.2	100
300	70.4	97.4	71.6	100

Table 3: Performance of the unsupervised normalization procedure. Only 1-best for each word was considered.

Once we obtained the 50 nearest neighbors for each word in the clean vocabulary, we used a combination of cosine metric threshold and Levenshtein distance (weighted equally) between the consonant

skeleton of the strings to construct the mapping lexicon. Finally, we inverted the table to obtain a normalization lexicon. Our procedure currently finds only one-to-one mappings. We took 60 singleton entries from the static normalization tables reported in Section 5.2 and evaluated the performance of our approach. The results are shown in Table 3 and some examples of learned normalizations are shown in Table 2.

5.3 Phrase Segmentation

In many SMS messages, multiple clauses may be concatenated without explicit punctuation. For example, the message *hi babe hope you're well sorry i missed your call* needs to be interpreted as *hi babe. hope you're well. sorry, i missed your call.* We perform phrase segmentation using an automatic punctuation classifier trained on SMS messages with punctuation. The classifier learns how to detect end of sentence markers, i.e. periods, as well as commas in the input stream of unpunctuated words.

An English punctuation classifier and a Spanish punctuation classifier was trained. The former was trained on two million words of smartphone data described in Section 4.1 while the latter was trained on 223,000 words of Spanish subtitles from the OpenSubtitles[5] corpus. From each of these data sets, a maximum entropy classifier was trained. Both classifiers utilized both unigram word and part of speech (POS) features of a window size of two words around the target word to be classified. A POS tagger trained on the English Penn Treebank provided English POS tags. Likewise, a Spanish POS tagger provided Spanish POS tags. The training data for the Spanish tagger, 1.6 million words in size, was obtained by running the Spanish Freeling parser over the Spanish version of TED talk transcripts. Results are shown in Table 4. Both phrase segmenters detect end of sentence well. The Spanish phrase segmenter detects commas better than the English one. This might be due to differences in the training sets; commas appear about 20 times more often in the Spanish data than in the English data.

	Class	Precision	Recall	F-measure
English	*period*	89.7	90.9	90.3
	comma	61.1	10.9	18.5
Spanish	*period*	94.3	87.4	90.7
	comma	74.2	37.4	49.7

Table 4: Performance of automatic phrase segmentation (numbers are in %)

5.4 Machine Translation

We used a phrase-based translation framework with the phrase table represented as a finite-state transducer (Rangarajan Sridhar et al., 2013). Our framework proceeds by using the standard procedure of performing word alignment using GIZA++ (Och and Ney, 2003) and obtaining phrases from the word alignment using heuristics (Zens and Ney, 2004) and subsequently scoring them. The phrase table is then represented as a finite-state transducer (FST). The FST decoder was used with minimum error rate training (MERT) to compute a set of weights for the log-linear model. It is important to note that the cost of arcs of the FST is a composite score (dot product of scores and weights) and hence requires an additional lookup during the N-best generation phase in MERT to obtain the component scores. The model is equivalent to Moses (**?**) phrase translation without reordering.

We noticed from the data collected in Section 4 that in typical SMS scenarios, a lot of phrases are stock phrases and hence caching these phrases may result in high accuracies instead of deriving the translation using a statistical model. We took the data created in Section 4 and created a FST to represent the sentences. The motivation is to increase the precision of common entries as well as reduce the latency involved in retrieving a translation from a statistical model. An example of the FST translation paradigm is shown in Figure 1

We experimented with the notion of using a consensus-based word alignment by combining the alignment obtained through different alignment tools. We used GIZA++ (Och and Ney, 2003), Berkeley

[5]http://www.opensubtitles.org

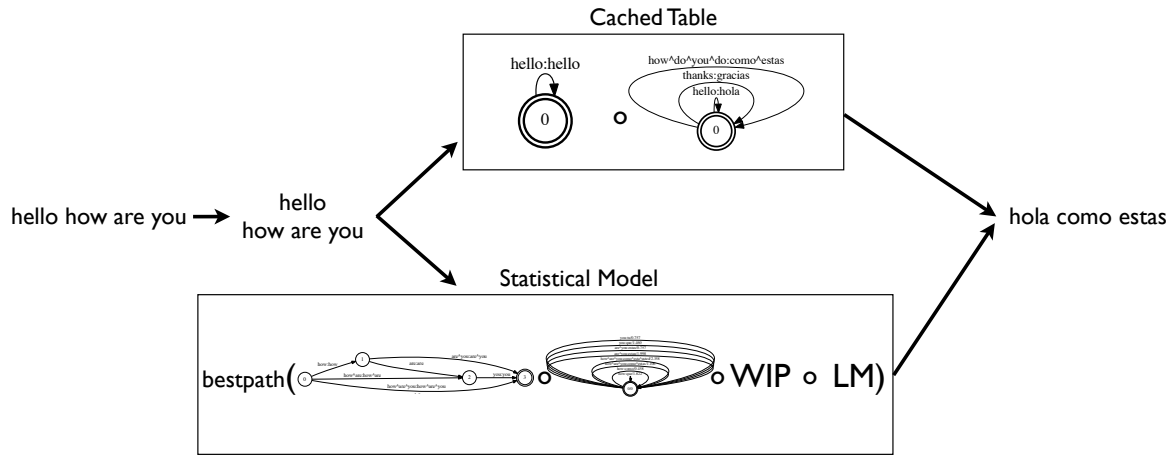

Figure 1: Illustration of the hybrid translation approach using FSTs. *WIP* and LM refer to the finite state automata for word insertion penalty and language model, respectively.

Alignment strategy	en2es	es2en
GIZA++	28.45	31.83
Pialign	28.08	33.48
Berkeley aligner	27.82	32.01
Union	28.01	33.14
Majority voting	27.32	32.96

Table 5: BLEU scores obtained using different alignment strategies. Only the statistical translation model was used in the evaluation.

aligner (Liang et al., 2006) and the Phrasal ITG aligner (Pialign) (Neubig et al., 2011). We combined the alignments in two different ways, taking the union of alignments or majority vote for each target word. For training the translation model, we used a total of 28.5 million parallel sentences obtained from the following sources: Opensubtitles (Tiedemann and Lars Nygaard, 2004), Europarl (Koehn, 2005), TED talks (Cettolo et al., 2012) and Web. The bitext was processed to eliminate spurious pairs by restricting the English and Spanish vocabularies to the top 150k frequent words as evidenced in a large collection of monolingual corpora. We also eliminated bitext with ratio of English to Spanish words less than 0.5. The initial model was optimized using MERT over 1000 parallel sentences from the SMS domain. Results of the machine translation experiments are shown in Table 5. The test set used was 456 messages collected in a real SMS interaction (see Section 6.1). The results indicate that consensus alignment procedure is not superior to the individual alignment outputs. Furthermore, the BLEU scores obtained through both the consensus procedures are not statistically significant with respect to the BLEU score obtained from the individual alignment tools. Hence, we used with the phrase translation table obtained using the Phrasal ITG aligner in all our experiments.

6 SMS Translation Service

In order to test the SMS translation models described in the previous sections, we created the infrastructure to intercept SMS messages, translate and deliver them in the preferred language of the recipient. The users were simply asked to register their numbers with a particular language through a Web portal and subsequently, all messages received by a user would be in the registered language. Some screenshots of interaction between users is shown in Figure 2. For the messages that are translated, we show both the original and translated messages. In cases where the translated message is longer than the character limit per message, we split the message over two message boxes.

6.1 User Evaluation

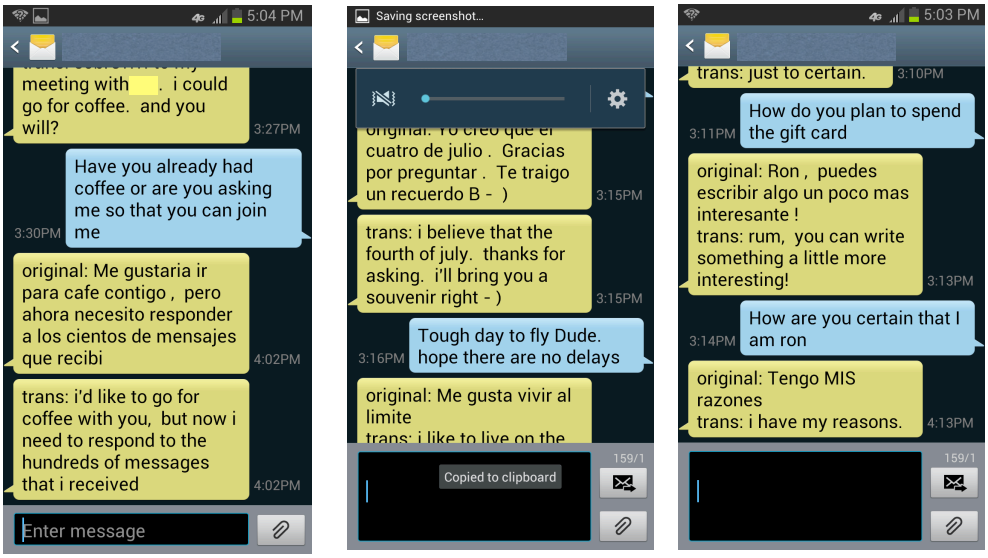

Figure 2: Screenshots of the SMS interface with translation

In order to test the SMS translation models described in the previous sections, we created the infrastructure to intercept SMS messages, translate and deliver them in the preferred language of the recipient. For the messages that are translated, we show both the original and translated messages. In cases where the translated message is longer than the character limit per message, we split the message over two message boxes. As part of the study we enrolled 20 English and 5 Spanish participants. The Spanish participants were bilingual while the English users had little to no knowledge of Spanish. Some of these interactions turned out to be short while others were had a large number of turns. We collected the messages exchanged over 2 days that amounted to 241 English and 215 Spanish messages.

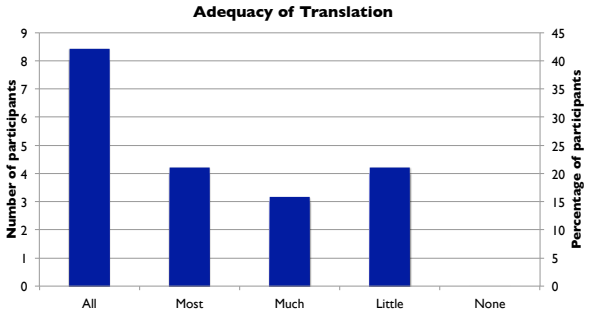

Figure 3: Subjective ratings regarding the adequacy of using SMS translation

We manually translated the 456 messages to create a test data set for evaluation purposes. In the absence of real SMS feeds in training, this test set is the closest we have to real SMS field data. The BLEU scores using the entire pipeline (normalization, punctuation, cached and statistical machine translation) for English-Spanish and Spanish-English was 31.25 and 37.19, respectively. We also created a survey for the participants to evaluate fluency and adequacy (LDC, 2005) Figures 3 and 4 show the survey results for adequacy and fluency, respectively. The results indicate that a majority of the people found the translation quality to be sufficiently adequate while the fluency was between *good* and *non-native*.

7 Discussion

The SMS bitext described in Section 4 consists of a total 58790 unique parallel sentences in the SMS domain. While the bulk of the data (speech-based) does not contain abbreviations and spelling errors, it

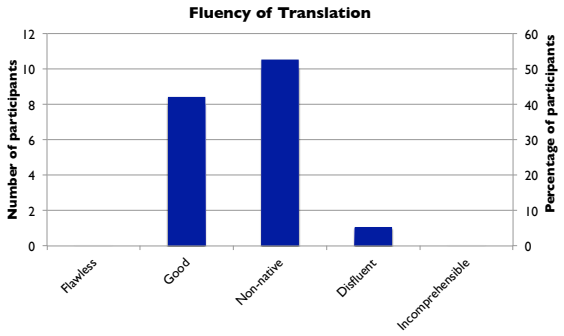

Figure 4: Subjective ratings regarding the fluency of using SMS translation

is highly representative of SMS messages and in fact is perfectly suited for statistical machine translation that typically uses normalized and tokenized data. The iterative procedure using Amazon Mechanical Turk is a good approach to procuring surrogate SMS data. We plan to continue harvesting data using this approach.

The unsupervised normalization lexicon learning using deep learning performs a good job of learning SMS shorthands. However, the induced lexicon contains only one-to-one word mappings. If one were to form compound words for a given dataset, the procedure can be potentially used for learning many-to-one and many-to-many mappings. Our framework also learns spelling errors rather well. It may also be possible to use distributed representations learned through log-linear models (Mikolov et al., 2013) for our task. However, this is beyond the scope of the work presented in this paper. Finally, we used only 1-best match for the unsupervised lexicon used in this work. One can potentially use a confusion network and compose it with the FST model to achieve higher accuracies. Our scheme results in fairly high precision with almost no false negatives (recall is extremely high) and can be reliably applied for normalization. The unsupervised normalization scheme did not yield significant improvements in BLEU score since our test set contained only 4 instances where shorthands were used.

Conventionally, sentence segmentation has been useful in improving the quality of statistical machine translation (Matusov et al., 2006; Matusov et al., 2005). Such segmentation, albeit into shorter phrases, is also useful for SMS translation. In the absence of phrase segmentation, the BLEU scores for English-Spanish and Spanish-English drop to 29.65 and 23.95, respectively. The degradation for Spanish-English messages is quite severe (drop from 37.19 to 23.95) as the lack of segmentation greatly reduces the use of the cached table. In the absence of segmentation, the cached table was used for 12.8% and 14.4% of the total phrases for English-Spanish and Spanish-English, respectively. However, with phrase segmentation the cached table was used for 29.2% and 39.2% of total phrases.

The subjective results obtained from the user trial augur well for the real use of translation technology as a feature in SMS. One of the issues in the study was balancing the English and Spanish participants. Since we had access to more English participants (20) in comparison with Spanish participants (5), the rate of exchange was slow. However, since SMS messages are not required to be real-time, participants still engaged in a meaningful conversation. Subjective evaluation results using LDC criteria indicate that most users were happy with the adequacy of translation while the fluency was rated as average. In general, SMS messages are not very fluent due to character limit imposed on the exchanges and hence machine translation has to use potentially disfluent source text.

8 Conclusion

We presented an application of statistical machine translation for translating SMS messages. We decoupled SMS translation into normalization followed by translation. Our unsupervised SMS normalization approach exploits the distributional similarity of words and learns SMS shorthands with good accuracy. We used a hybrid translation approach to exploit the repetitive nature of high frequency SMS messages. Both objective and subjective evaluation experiments indicate that our system generates translation with high quality while addressing the idiosyncrasies of SMS messages.

References

A. Aw, M. Zhang, J. Xiao, and J. Su. 2009. A phrase-based statistical model for SMS text normalization. In *Proceedings of COLING*, pages 33–40.

S. Bangalore, V. Murdock, and G. Riccardi. 2002. Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *Proceedings of COLING*.

R. Beaufort, S. Roekhaut, L. A. Cougnon, and C. Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing sms messages. In *Proceedings of ACL*, pages 770–779.

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

M. Cettolo, C. Girardi, and M. Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of EAMT*.

T. Chen and M. Y. Kan. 2013. Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation*, 47(2):299–335.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML*.

R. Collobert, K. Kavukcuoglu, and C. Farabet. 2011. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*.

V. Eidelman, K. Hollingshead, and P. Resnik. 2011. Noisy SMS Machine Translation in Low-Density Languages. In *Proceedings of 6th Workshop on Statistical Machine Translation*.

C. Fairon and S. Paumier. 2006. A translated corpus of 30,000 french SMS. In *Proceedings of LREC*.

C. Kobus, F. Yvon, and G. Damnati. 2008. Normalizing sms: Are two metaphors better than one? In *Proceedings of COLING*, pages 441–448.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report, Revision 1.5.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of NAACL-HLT*, pages 104–111.

E. Matusov, G. Leusch, O. Bender, and H. Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of IWSLT*, pages 148–154.

E. Matusov, A. Mauser, and H. Ney. 2006. Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proceedings of IWSLT*, pages 158–165.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the ACL*.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

D. Pennell and Y. Liu. 2011. A character-level machine translation approach for normalization of SMS abbreviations. In *Proceedings of IJCNLP*.

V. K. Rangarajan Sridhar, J. Chen, S. Bangalore, A. Ljolje, and R. Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proceedings of NAACL-HLT*.

E. Sanders. 2012. Collecting and analysing chats and tweets in SoNaR. In *Proceedings of LREC*.

C. Tagg. 2009. *Across-frequency in convolutive blind source separation*. dissertation, University of Birmingham.

J. Tiedemann and L. Lars Nygaard. 2004. The OPUS corpus - parallel & free. In *Proceedings of LREC*.

M. Treurniet, O. De Clercq, H. van den Heuvel, and N. Oostdijk. 2012. Collecting a corpus of Dutch SMS. In *Proceedings of LREC*, pages 2268–2273.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL*.

P. Wang and H. Tou Ng. 2013. A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of NAACL-HLT*.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *In Proceedings of HLT-NAACL*, pages 257–264.