A Phrase-based Statistical Model for SMS Text Normalization

AiTi Aw, Min Zhang, Juan Xiao, Jian Su
Institute of Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
{aaiti,mzhang,stuxj,sujian}@i2r.a-star.edu.sg

Abstract

Short Messaging Service (SMS) texts behave quite differently from normal written texts and have some very special phenomena. To translate SMS texts, traditional approaches model such irregularities directly in Machine Translation (MT). However, such approaches suffer from customization problem as tremendous effort is required to adapt the language model of the existing translation system to handle SMS text style. We offer an alternative approach to resolve such irregularities by normalizing SMS texts before MT. In this paper, we view the task of SMS normalization as a translation problem from the SMS language to the English language [1] and we propose to adapt a phrase-based statistical MT model for the task. Evaluation by 5-fold cross validation on a parallel SMS normalized corpus of 5000 sentences shows that our method can achieve 0.80702 in BLEU score against the baseline BLEU score 0.6958. Another experiment of translating SMS texts from English to Chinese on a separate SMS text corpus shows that, using SMS normalization as MT preprocessing can largely boost SMS translation performance from 0.1926 to 0.3770 in BLEU score.

1 Motivation

SMS translation is a mobile Machine Translation (MT) application that translates a message from one language to another. Though there exists many commercial MT systems, direct use of such systems fails to work well due to the special phenomena in SMS texts, e.g. the unique relaxed and creative writing style and the frequent use of unconventional and not yet standardized short-forms. Direct modeling of these special phenomena in MT requires tremendous effort. Alternatively, we can normalize SMS texts into grammatical texts before MT. In this way, the traditional MT is treated as a "black-box" with little or minimal adaptation. One advantage of this pre-translation normalization is that the diversity in different user groups and domains can be modeled separately without accessing and adapting the language model of the MT system for each SMS application. Another advantage is that the normalization module can be easily utilized by other applications, such as SMS to voicemail and SMS-based information query.

In this paper, we present a phrase-based statistical model for SMS text normalization. The normalization is visualized as a translation problem where messages in the SMS language are to be translated to normal English using a similar phrase-based statistical MT method (Koehn et al., 2003). We use IBM's BLEU score (Papineni et al., 2002) to measure the performance of SMS text normalization. BLEU score computes the similarity between two sentences using n-gram statistics, which is widely-used in MT evaluation. A set of parallel SMS messages, consisting of 5000 raw (un-normalized) SMS messages and their manually normalized references, is constructed for training and testing. Evaluation by 5-fold cross validation on this corpus shows that our method can achieve accuracy of 0.80702 in BLEU score compared to the baseline system of 0.6985. We also study the impact of our SMS text normalization on the task of SMS translation. The experiment of translating SMS texts from English to Chinese on a corpus comprising 402 SMS texts shows that, SMS normalization as a preprocessing step of MT can boost the translation performance from 0.1926 to 0.3770 in BLEU score.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 summarizes the characteristics of English SMS texts. Section 4 discusses our method and Section 5 reports our experiments. Section 6 concludes the paper.

2 Related Work

There is little work reported on SMS normalization and translation. Bangalore et al. (2002) used

[1] This paper only discusses English SMS text normalization.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 33–40,
Sydney, July 2006.

a consensus translation technique to bootstrap parallel data using off-the-shelf translation systems for training a hierarchical statistical translation model for general domain instant messaging used in Internet chat rooms. Their method deals with the special phenomena of the instant messaging language (rather than the SMS language) in each individual MT system. Clark (2003) proposed to unify the process of tokenization, segmentation and spelling correction for normalization of general noisy text (rather than SMS or instant messaging texts) based on a noisy channel model at the character level. However, results of the normalization are not reported. Aw et al. (2005) gave a brief description on their input pre-processing work for an English-to-Chinese SMS translation system using a word-group model. In addition, in most of the commercial SMS translation applications [2], SMS lingo (i.e., SMS short form) dictionary is provided to replace SMS short-forms with normal English words. Most of the systems do not handle OOV (out-of-vocabulary) items and ambiguous inputs. Following compares SMS text normalization with other similar or related applications.

2.1 SMS Normalization versus General Text Normalization

General text normalization deals with Non-Standard Words (NSWs) and has been well-studied in text-to-speech (Sproat et al., 2001) while SMS normalization deals with Non-Words (NSs) or lingoes and has seldom been studied before. NSWs, such as digit sequences, acronyms, mixed case words (WinNT, SunOS), abbreviations and so on, are grammatically correct in linguistics. However lingoes, such as "*b4*" (*before*) and "*bf*" (*boyfriend*), which are usually self-created and only accepted by young SMS users, are not yet formalized in linguistics. Therefore, the special phenomena in SMS texts impose a big challenge to SMS normalization.

2.2 SMS Normalization versus Spelling Correction Problem

Intuitively, many would regard SMS normalization as a spelling correction problem where the lingoes are erroneous words or non-words to be replaced by English words. Researches on spelling correction centralize on typographic and cognitive/orthographic errors (Kukich, 1992) and use approaches (M.D. Kernighan, Church and Gale, 1991) that mostly model the edit operations using distance measures (Damerau 1964; Levenshtein 1966), specific word set confusions (Golding and Roth, 1999) and pronunciation modeling (Brill and Moore, 2000; Toutanova and Moore, 2002). These models are mostly character-based or string-based without considering the context. In addition, the author might not be aware of the errors in the word introduced during the edit operations, as most errors are due to mistype of characters near to each other on the keyboard or homophones, such as "poor" or "pour".

In SMS, errors are not isolated within word and are usually not surrounded by clean context. Words are altered deliberately to reflect sender's distinct creation and idiosyncrasies. A character can be deleted on purpose, such as "*wat*" (*what*) and "*hv"* (*have*). It also consists of short-forms such as "*b4*" (*before*), "*bf*" (*boyfriend*). In addition, normalizing SMS text might require the context to be spanned over more than one lexical unit such as "*lemme*" (*let me*), "*ur*" (*you are*) etc. Therefore, the models used in spelling correction are inadequate for providing a complete solution for SMS normalization.

2.3 SMS Normalization versus Text Paraphrasing Problem

Others may regard SMS normalization as a paraphrasing problem. Broadly speaking, paraphrases capture core aspects of variability in language, by representing equivalencies between different expressions that correspond to the same meaning. In most of the recent works (Barzilay and McKeown, 2001; Shimohata, 2002), they are acquired (semi-) automatically from large comparable or parallel corpora using lexical and morpho-syntactic information.

Text paraphrasing works on clean texts in which contextual and lexical-syntactic features can be extracted and used to find "approximate conceptual equivalence". In SMS normalization, we are dealing with non-words and "ungrammatically" sentences with the purpose to normalize or standardize these words and form better sentences. The SMS normalization problem is thus different from text paraphrasing. On the other hand, it bears some similarities with MT as we are trying to "convert" text from one language to another. However, it is a simpler problem as most of the time; we can find the same word in both the source and target text, making alignment easier.

[2] http://www.etranslator.ro and http://www.transl8bit.com

3 Characteristics of English SMS

Our corpus consists of 55,000 messages collected from two sources, a SMS chat room and correspondences between university students. The content is mostly related to football matches, making friends and casual conversations on "how, what and where about". We summarize the text behaviors into two categories as below.

3.1 Orthographic Variation

The most significant orthographic variant in SMS texts is in the use of non-standard, self-created short-forms. Usually, sender takes advantage of phonetic spellings, initial letters or number homophones to mimic spoken conversation or shorten words or phrases (*hw* vs. *homework* or *how*, *b4* vs. *before*, *cu* vs. *see you*, *2u* vs. *to you*, *oic* vs. *oh I see*, etc.) in the attempt to minimize key strokes. In addition, senders create a new form of written representation to express their oral utterances. Emotions, such as ":(" symbolizing sad, ":)" symbolizing smiling, ":()" symbolizing shocked, are representations of body language. Verbal effects such as "*hehe*" for laughter and emphatic discourse particles such as "*lor*", "*lah*", "*meh*" for colloquial English are prevalent in the text collection.

The loss of "alpha-case" information posts another challenge in lexical disambiguation and introduces difficulty in identifying sentence boundaries, proper nouns, and acronyms. With the flexible use of punctuation or not using punctuation at all, translation of SMS messages without prior processing is even more difficult.

3.2 Grammar Variation

SMS messages are short, concise and convey much information within the limited space quota (160 letters for English), thus they tend to be implicit and influenced by pragmatic and situation reasons. These inadequacies of language expression such as deletion of articles and subject pronoun, as well as problems in number agreements or tenses make SMS normalization more challenging. Table 1 illustrates some orthographic and grammar variations of SMS texts.

3.3 Corpus Statistics

We investigate the corpus to assess the feasibility of replacing the lingoes with normal English words and performing limited adjustment to the text structure. Similarly to Aw et al. (2005), we focus on the three major cases of transformation as shown in the corpus: (1) replacement of OOV words and non-standard SMS lingoes; (2) removal of slang and (3) insertion of auxiliary or copula verb and subject pronoun.

Phenomena	Messages
1. Dropping '?' at the end of question	*btw, wat is ur view* (*By the way, what is your view?*)
2. Not using any punctuation at all	*Eh speak english mi malay not tt good* (*Eh, speak English! My Malay is not that good.*)
3. Using spelling/punctuation for emphasis	*goooooood Sunday morning !!!!!!* (*Good Sunday morning!*)
4. Using phonetic spelling	*dat iz enuf* (*That is enough*)
5. Dropping vowel	*i hv cm to c my luv.* (*I have come to see my love.*)
6. Introducing local flavor	*yar lor where u go juz now* (*yes, where did you go just now?*)
7. Dropping verb	*I hv 2 go. Dinner w parents.* (*I have to go. Have dinner with parents.*)

Table 1. Examples of SMS Messages

Transformation	Percentage (%)
Insertion	8.09
Deletion	5.48
Substitution	86.43

Table 2. Distribution of Insertion, Deletion and Substitution Transformation.

Substitution	Deletion	Insertion
u -> you	*m*	*are*
2 → to	*lah*	*am*
n → and	*t*	*is*
r → are	*ah*	*you*
ur →your	*leh*	*to*
dun → don't	*1*	*do*
man → manchester	*huh*	*a*
no → number	*one*	*in*
intro → introduce	*lor*	*yourself*
wat → what	*ahh*	*will*

Table 3. Top 10 Most Common Substitution, Deletion and Insertion

Table 2 shows the statistics of these transformations based on 700 messages randomly selected, where 621 (88.71%) messages required

normalization with a total of 2300 transformations. Substitution accounts for almost 86% of all transformations. Deletion and substitution make up the rest. Table 3 shows the top 10 most common transformations.

4 SMS Normalization

We view the SMS language as a variant of English language with some derivations in vocabulary and grammar. Therefore, we can treat SMS normalization as a MT problem where the SMS language is to be translated to normal English. We thus propose to adapt the statistical machine translation model (Brown et al., 1993; Zens and Ney, 2004) for SMS text normalization. In this section, we discuss the three components of our method: modeling, training and decoding for SMS text normalization.

4.1 Basic Word-based Model

The SMS normalization model is based on the source channel model (Shannon, 1948). Assuming that an English sentence *e*, of length N is "corrupted" by a noisy channel to produce a SMS message *s*, of length M, the English sentence *e*, could be recovered through a posteriori distribution for a channel target text given the source text $P(s \mid e)$, and a prior distribution for the channel source text $P(e)$.

$$\begin{aligned} \hat{e}_1^N &= \arg\max_{e_1^N} \left\{ P(e_1^N \mid s_1^M) \right\} \\ &= \arg\max_{e_1^N} \left\{ P(s_1^M \mid e_1^N) \bullet P(e_1^N) \right\} \end{aligned} \tag{1}$$

Assuming that one SMS word is mapped exactly to one English word in the channel model $P(s \mid e)$ under an alignment A, we need to consider only two types of probabilities: the alignment probabilities denoted by $P(m \mid a_m)$ and the lexicon mapping probabilities denoted by $P(s_m \mid e_{a_m})$ (Brown et al. 1993). The channel model can be written as in the following equation where *m* is the position of a word in *s* and a_m its alignment in e.

$$\begin{aligned} P(s_1^M \mid e_1^N) &= \sum_A P(s_1^M, A \mid e_1^N) \\ &= \sum_A P(A \mid e_1^N) \bullet P(s_1^M \mid A, e_1^N) \\ &\approx \sum_A \left(\prod_{m=1}^{M} \left\{ P(m \mid a_m) \bullet P(s_m \mid e_{a_m}) \right\} \right) \end{aligned} \tag{2}$$

If we include the word "null" in the English vocabulary, the above model can fully address the deletion and substitution transformations, but inadequate to address the insertion transformation. For example, the lingoes *"duno"*, *"ysnite"* have to be normalized using an insertion transformation to become *"don't know"* and *"yesterday night"*. Moreover, we also want the normalization to have better lexical affinity and linguistic equivalent, thus we extend the model to allow many words to many words alignment, allowing a sequence of SMS words to be normalized to a sequence of contiguous English words. We call this updated model a phrase-based normalization model.

4.2 Phrase-based Model

Given an English sentence *e* and SMS sentence *s*, if we assume that *e* can be decomposed into K phrases with a segmentation T, such that each phrase $\tilde{e}_k$ in *e* can be corresponded with one phrase $\tilde{s}_k$ in *s*, we have $e_1^N = \tilde{e}_1 \ldots \tilde{e}_k \ldots \tilde{e}_K$ and $s_1^M = \tilde{s}_1 \ldots \tilde{s}_k \ldots \tilde{s}_K$. The channel model can be rewritten in equation (3).

$$\begin{aligned} P(s_1^M \mid e_1^N) &= \sum_T P(s_1^M, T \mid e_1^N) \\ &= \sum_T P(T \mid e_1^N) \bullet P(s_1^M \mid T, e_1^N) \\ &= \sum_T P(T \mid e_1^N) \bullet P(\tilde{s}_1^K \mid \tilde{e}_1^K) \\ &\approx \max_T \left\{ P(T \mid e_1^N) \bullet P(\tilde{s}_1^K \mid \tilde{e}_1^K) \right\} \end{aligned} \tag{3}$$

This is the basic function of the channel model for the phrase-based SMS normalization model, where we used the maximum approximation for the sum over all segmentations. Then we further decompose the probability $P(\tilde{s}_1^K \mid \tilde{e}_1^K)$ using a phrase alignment $\tilde{A}$ as done in the previous word-based model.

$$\begin{aligned} P(\tilde{s}_1^K \mid \tilde{e}_1^K) &= \sum_{\tilde{A}} P(\tilde{s}_1^K, \tilde{A} \mid \tilde{e}_1^K) \\ &= \sum_{\tilde{A}} \left\{ P(\tilde{A} \mid \tilde{e}_1^K) \bullet P(\tilde{s}_1^K \mid \tilde{A}, \tilde{e}_1^K) \right\} \\ &= \sum_{\tilde{A}} \left(\prod_{k=1}^{K} \left\{ P(k \mid \tilde{a}_k) \bullet P(\tilde{s}_k \mid \tilde{s}_1^{k-1}, \tilde{e}_{\tilde{a}_1}^{\tilde{a}_k}) \right\} \right) \\ &\approx \sum_{\tilde{A}} \left(\prod_{k=1}^{K} \left\{ P(k \mid \tilde{a}_k) \bullet P(\tilde{s}_k \mid \tilde{e}_{\tilde{a}_k}) \right\} \right) \end{aligned} \tag{4}$$

We are now able to model the three transformations through the normalization pair $(\tilde{s}_k, \tilde{e}_{\tilde{a}_k})$,

with the mapping probability $P(\tilde{s}_k \mid \tilde{e}_{\tilde{a}_k})$. The followings show the scenarios in which the three transformations occur.

Insertion	$\lvert\tilde{s}_k\rvert < \lvert\tilde{e}_{\tilde{a}_k}\rvert$
Deletion	$\tilde{e}_{\tilde{a}_k}$ = null
Substitution	$\lvert\tilde{s}_k\rvert = \lvert\tilde{e}_{\tilde{a}_k}\rvert$

The statistics in our training corpus shows that by selecting appropriate phrase segmentation, the position re-ordering at the phrase level occurs rarely. It is not surprising since most of the English words or phrases in normal English text are replaced with lingoes in SMS messages without position change to make SMS text short and concise and to retain the meaning. Thus we need to consider only monotone alignment at phrase level, i.e., $k = \tilde{a}_k$, as in equation (4). In addition, the word-level reordering within phrase is learned during training. Now we can further derive equation (4) as follows:

$$\begin{aligned} P(\tilde{s}_1^K \mid \tilde{e}_1^K) &\approx \sum_{\tilde{A}} \left(\prod_{k=1}^{K} \left\{ P(k \mid \tilde{a}_k) \bullet P(\tilde{s}_k \mid \tilde{e}_{\tilde{a}_k}) \right\} \right) \\ &\approx \prod_{k=1}^{K} P(\tilde{s}_k \mid \tilde{e}_k) \end{aligned} \quad (5)$$

The mapping probability $P(\tilde{s}_k \mid \tilde{e}_k)$ is estimated via relative frequencies as follows:

$$P(\tilde{s}_k \mid \tilde{e}_k) = \frac{N(\tilde{s}_k, \tilde{e}_k)}{\sum_{\tilde{s}_k'} N(\tilde{s}_k', \tilde{e}_k)} \quad (6)$$

Here, $N(\tilde{s}_k, \tilde{e}_k)$ denotes the frequency of the normalization pair $(\tilde{s}_k, \tilde{e}_k)$.

Using a bigram language model and assuming Bayes decision rule, we finally obtain the following search criterion for equation (1).

$$\begin{aligned} \hat{e}_1^N &= \arg\max_{e_1^N} \left\{ P(e_1^N) \bullet P(s_1^M \mid e_1^N) \right\} \\ &\approx \arg\max_{e_1^N} \left\{ \prod_{n=1}^{N} P(e_n \mid e_{n-1}) \right. \\ &\qquad \left. \bullet \max_T \left\{ P(T \mid e_1^N) \bullet \prod_{k=1}^{K} P(\tilde{s}_k \mid \tilde{e}_k) \right\} \right\} \\ &\approx \arg\max_{e_1^N, T} \left\{ \prod_{n=1}^{N} P(e_n \mid e_{n-1}) \bullet \prod_{k=1}^{K} P(\tilde{s}_k \mid \tilde{e}_k) \right\} \end{aligned} \quad (7)$$

For the above equation, we assume the segmentation probability $P(T \mid e_1^N)$ to be constant. Finally, the SMS normalization model consists of two sub-models: **a word-based language model** (LM), characterized by $P(e_n \mid e_{n-1})$ and **a phrase-based lexical mapping model** (channel model), characterized by $P(\tilde{s}_k \mid \tilde{e}_k)$.

4.3 Training Issues

For the phrase-based model training, the sentence-aligned SMS corpus needs to be aligned first at the phrase level. The maximum likelihood approach, through EM algorithm and Viterbi search (Dempster et al., 1977) is employed to infer such an alignment. Here, we make a reasonable assumption on the alignment unit that a single SMS word can be mapped to a sequence of contiguous English words, but not vice verse. The EM algorithm for phrase alignment is illustrated in Figure 1 and is formulated by equation (8).

The Expectation-Maximization Algorithm

(1) Bootstrap initial alignment using orthographic similarities
(2) Expectation: Update the joint probabilities $P(\tilde{s}_k, \tilde{e}_k)$
(3) Maximization: Apply the joint probabilities $P(\tilde{s}_k, \tilde{e}_k)$ to get new alignment using Viterbi search algorithm
(4) Repeat (2) to (3) until alignment converges
(5) Derive normalization pairs from final alignment

Figure 1. Phrase Alignment Using EM Algorithm

$$\hat{\gamma}_{<\tilde{s}_k, \tilde{e}_k>} = \arg\max_{\hat{\gamma}_{<\tilde{s}_k, \tilde{e}_k>}} \prod_{k=1}^{K} P(\tilde{s}_k, \tilde{e}_k \mid s_1^M, e_1^N) \quad (8)$$

The alignment process given in equation (8) is different from that of normalization given in equation (7) in that, here we have an aligned input sentence pair, s_1^M and e_1^N. The alignment process is just to find the alignment segmentation $\hat{\gamma}_{<\tilde{s}_k, \tilde{e}_k>} = <\tilde{s}_k, \tilde{e}_k>_{k=1,K}$ between the two sentences that maximizes the joint probability. Therefore, in step (2) of the EM algorithm given at Figure 1, only the joint probabilities $P(\tilde{s}_k, \tilde{e}_k)$ are involved and updated.

Since EM may fall into local optimization, in order to speed up convergence and find a nearly global optimization, a string matching technique is exploited at the initialization step to identify the most probable normalization pairs. The or-

thographic similarities captured by edit distance and a SMS lingo dictionary[3] which contains the commonly used short-forms are first used to establish phrase mapping boundary candidates. Heuristics are then exploited to match tokens within the pairs of boundary candidates by trying to combine consecutive tokens within the boundary candidates if the numbers of tokens do not agree.

Finally, a filtering process is carried out to manually remove the low-frequency noisy alignment pairs. Table 4 shows some of the extracted normalization pairs. As can be seen from the table, our algorithm discovers ambiguous mappings automatically that are otherwise missing from most of the lingo dictionary.

$(\tilde{s},\tilde{e})$	$\log P(\tilde{s}\mid\tilde{e})$
(2, 2)	0
(2, to)	-0.579466
(2, too)	-0.897016
(2, null)	-2.97058
(4, 4)	0
(4, for)	-0.431364
(4, null)	-3.27161
(w, who are)	-0.477121
(w, with)	-0.764065
(w, who)	-1.83885
(dat, that)	-0.726999
(dat, date)	-0.845098
(tmr, tomorrow)	-0.341514

Table 4. Examples of normalization pairs

Given the phrase-aligned SMS corpus, the lexical mapping model, characterized by $P(\tilde{s}_k \mid \tilde{e}_k)$, is easily to be trained using equation (6). Our n-gram LM $P(e_n \mid e_{n-1})$ is trained on English Gigaword provided by LDC using SRILM language modeling toolkit (Stolcke, 2002). Backoff smoothing (Jelinek, 1991) is used to adjust and assign a non-zero probability to the unseen words to address data sparseness.

4.4 Monotone Search

Given an input s, the search, characterized in equation (7), is to find a sentence e that maximizes $P(s \mid e)\bullet P(e)$ using the normalization model. In this paper, the maximization problem in equation (7) is solved using a monotone search, implemented as a Viterbi search through dynamic programming.

5 Experiments

The aim of our experiment is to verify the effectiveness of the proposed statistical model for SMS normalization and the impact of SMS normalization on MT.

A set of 5000 parallel SMS messages, which consists of raw (un-normalized) SMS messages and reference messages manually prepared by two project members with inter-normalization agreement checked, was prepared for training and testing. For evaluation, we use IBM's BLEU score (Papineni et al., 2002) to measure the performance of the SMS normalization. BLEU score measures the similarity between two sentences using n-gram statistics with a penalty for too short sentences, which is already widely-used in MT evaluation.

Setup	**BLEU score (3-gram)**
Raw SMS without Normalization	0.5784
Dictionary Look-up plus Frequency	0.6958
Bi-gram Language Model Only	0.7086

Table 5. Performance of different setups of the baseline experiments on the 5000 parallel SMS messages

5.1 Baseline Experiments: Simple SMS Lingo Dictionary Look-up and Using Language Model Only

The baseline experiment is to moderate the texts using a lingo dictionary comprises 142 normalization pairs, which is also used in bootstrapping the phrase alignment learning process.

Table 5 compares the performance of the different setups of the baseline experiments. We first measure the complexity of the SMS normalization task by directly computing the similarity between the raw SMS text and the normalized English text. The 1st row of Table 5 reports the similarity as 0.5784 in BLEU score, which implies that there are quite a number of English word 3-gram that are common in the raw and normalized messages. The 2nd experiment is carried out using only simple dictionary look-up.

[3] The entries are collected from various websites such as http://www.handphones.info/sms-dictionary/sms-lingo.php, and http://www.funsms.net/sms_dictionary.htm, etc.

Lexical ambiguity is addressed by selecting the highest-frequency normalization candidate, i.e., only unigram LM is used. The performance of the 2nd experiment is 0.6958 in BLEU score. It suggests that the lingo dictionary plus the unigram LM is very useful for SMS normalization. Finally we carry out the 3rd experiment using dictionary look-up plus bi-gram LM. Only a slight improvement of 0.0128 (0.7086-0.6958) is obtained. This is largely because the English words in the lingo dictionary are mostly high-frequency and commonly-used. Thus bi-gram does not show much more discriminative ability than unigram without the help of the phrase-based lexical mapping model.

5.2 Using Phrase-based Model

We then conducted the experiment using the proposed method (Bi-gram LM plus a phrase-based lexical mapping model) through a five-fold cross validation on the 5000 parallel SMS messages. Table 6 shows the results. An average score of 0.8070 is obtained. Compared with the baseline performance in Table 5, the improvement is very significant. It suggests that the phrase-based lexical mapping model is very useful and our method is effective for SMS text normalization. Figure 2 is the learning curve. It shows that our algorithm converges when training data is increased to 3000 SMS parallel messages. This suggests that our collected corpus is representative and enough for training our model. Table 7 illustrates some examples of the normalization results.

5-fold cross validation	BLEU score (3-gram)
Setup 1	0.8023
Setup 2	0.8236
Setup 3	0.8071
Setup 4	0.8113
Setup 5	0.7908
Ave.	0.8070

Table 6. Normalization results for 5-fold cross validation test

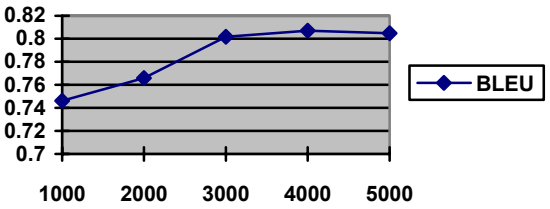

Figure 2. Learning Curve

Experimental result analysis reveals that the strength of our model is in its ability to disambiguate mapping as in "*2*" to "*two*" or "*to*" and "*w*" to "w*ith*" or "*who*". Error analysis shows that the challenge of the model lies in the proper insertion of subject pronoun and auxiliary or copula verb, which serves to give further semantic information about the main verb, however this requires significant context understanding. For example, a message such as "*u smart*" gives little clues on whether it should be normalized to "*Are you smart*?" or "*You are smart*." unless the full conversation is studied.

Takako w r u?
Takako who are you?
Im in ns, lik soccer, clubbin hangin w frenz! Wat bout u mee?
I'm in ns, like soccer, clubbing hanging with friends! What about you?
fancy getting excited w others' boredom
Fancy getting excited with others' boredom
If u ask me b4 he ask me then i'll go out w u all lor. N u still can act so real.
If you ask me before he asked me then I'll go out with you all. And you still can act so real.
Doing nothing, then u not having dinner w us?
Doing nothing, then you do not having dinner with us?
Aiyar sorry lor forgot 2 tell u... Mtg at 2 pm.
Sorry forgot to tell you... Meeting at two pm.
tat's y I said it's bad dat all e gals know u... Wat u doing now?
That's why I said it's bad that all the girls know you... What you doing now?

Table 7. Examples of Normalization Results

5.3 Effect on English-Chinese MT

An experiment was also conducted to study the effect of normalization on MT using 402 messages randomly selected from the text corpus. We compare three types of SMS message: raw SMS messages, normalized messages using simple dictionary look-up and normalized messages using our method. The messages are passed to two different English-to-Chinese translation systems provided by Systran[4] and Institute for Infocomm Research[5](I^{2}R) separately to produce three sets of translation output. The translation quality is measured using 3-gram cumulative BLEU score against two reference messages. 3-gram is

[4] http://www.systranet.com/systran/net
[5] http://nlp.i2r.a-star.edu.sg/techtransfer.html

used as most of the messages are short with average length of seven words. Table 8 shows the details of the BLEU scores. We obtain an average of 0.3770 BLEU score for normalized messages against 0.1926 for raw messages. The significant performance improvement suggests that preprocessing of normalizing SMS text using our method before MT is an effective way to adapt a general MT system to SMS domain.

	I^2R	**Systran**	**Ave.**
Raw Message	0.2633	0.1219	0.1926
Dict Lookup	0.3485	0.1690	0.2588
Normalization	0.4423	0.3116	0.3770

Table 8. SMS Translation BLEU score with or without SMS normalization

6 Conclusion

In this paper, we study the differences among SMS normalization, general text normalization, spelling check and text paraphrasing, and investigate the different phenomena of SMS messages. We propose a phrase-based statistical method to normalize SMS messages. The method produces messages that collate well with manually normalized messages, achieving 0.8070 BLEU score against 0.6958 baseline score. It also significantly improves SMS translation accuracy from 0.1926 to 0.3770 in BLEU score without adjusting the MT model.

This experiment results provide us with a good indication on the feasibility of using this method in performing the normalization task. We plan to extend the model to incorporate mechanism to handle missing punctuation (which potentially affect MT output and are not being taken care at the moment), and making use of pronunciation information to handle OOV caused by the use of phonetic spelling. A bigger data set will also be used to test the robustness of the system leading to a more accurate alignment and normalization.

References

A.T. Aw, M. Zhang, Z.Z. Fan, P.K. Yeo and J. Su. 2005. *Input Normalization for an English-to-Chinese SMS Translation System.* MT Summit-2005

S. Bangalore, V. Murdock and G. Riccardi. 2002. *Bootstrapping Bilingual Data using Consensus Translation for a Multilingual Instant Messaging System.* COLING-2002

R. Barzilay and K. R. McKeown. 2001. *Extracting paraphrases from a parallel corpus*. ACL-2001

E. Brill and R. C. Moore. 2000. *An Improved Error Model for Noisy Channel Spelling Correction.* ACL-2000

P. F. Brown, S. D. Pietra, V. D. Pietra and R. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics: 19(2)

A. Clark. 2003. *Pre-processing very noisy text.* In Proceedings of Workshop on Shallow Processing of Large Corpora, Lancaster, 2003

F. J. Damerau. 1964. *A technique for computer detection and correction of spelling errors*. Communications ACM 7, 171-176

A.P. Dempster, N.M. Laird and D.B. Rubin. 1977. *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Series B, Vol. 39, 1-38

A. Golding and D. Roth. 1999. *A Winnow-Based Approach to Spelling Correction*. Machine Learning 34: 107-130

F. Jelinek. 1991. *Self-organized language modeling for speech recognition*. In A. Waibel and K.F. Lee, editors, Readings in Speech Recognition, pages 450-506. Morgan Kaufmann, 1991

M. D. Kernighan, K Church and W. Gale. 1990. *A spelling correction program based on a noisy channel model.* COLING-1990

K. Kukich. 1992. *Techniques for automatically correcting words in text*. ACM Computing Surveys, 24(4):377-439

K. A. Papineni, S. Roukos, T. Ward and W. J. Zhu. 2002. *BLEU : a Method for Automatic Evaluation of Machine Translation*. ACL-2002

P. Koehn, F.J. Och and D. Marcu. 2003. *Statistical Phrase-Based Translation*. HLT-NAACL-2003

C. Shannon. 1948. *A mathematical theory of communication*. Bell System Technical Journal 27(3): 379-423

M. Shimohata and E. Sumita 2002. *Automatic Paraphrasing Based on Parallel Corpus for Normalization*. LREC-2002

R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf and C. Richards. 2001. *Normalization of Non-Standard Words*. Computer Speech and Language, 15(3):287-333

A. Stolcke. 2002. SRILM – *An extensible language modeling toolkit*. ICSLP-2002

K. Toutanova and R. C. Moore. 2002. *Pronunciation Modeling for Improved Spelling Correction.* ACL-2002

R. Zens and H. Ney. 2004. *Improvements in Phrase-Based Statistical MT*. HLT-NAALL-2004