A Beam-Search Decoder for Normalization of Social Media Text with Application to Machine Translation

Pidong Wang
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
wangpd@comp.nus.edu.sg

Hwee Tou Ng
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
nght@comp.nus.edu.sg

Abstract

Social media texts are written in an informal style, which hinders other natural language processing (NLP) applications such as machine translation. Text normalization is thus important for processing of social media text. Previous work mostly focused on normalizing words by replacing an informal word with its formal form. In this paper, to further improve other downstream NLP applications, we argue that other normalization operations should also be performed, e.g., missing word recovery and punctuation correction. A novel beam-search decoder is proposed to effectively integrate various normalization operations. Empirical results show that our system obtains statistically significant improvements over two strong baselines in both normalization and translation tasks, for both Chinese and English.

1 Introduction

Social media texts include SMS (Short Message Service) messages, Twitter messages, Facebook updates, etc. They are different from formal texts due to their significant informal characteristics, so they always pose difficulties for applications such as machine translation (MT) (Aw et al., 2005) and named entity recognition (Liu et al., 2011), because of a lack of training data containing informal texts. Thus, the applications always suffer from a substantial performance drop when evaluated on social media texts. For example, Ritter et al. (2011) reported a drop from 90% to 76% on part-of-speech tagging, and Foster et al. (2011) found a drop of 20% in dependency parsing.

Creating training data of social media texts specifically for a text processing task is time-consuming. For example, to create parallel Chinese-English training texts for translation of social media texts, it takes three minutes on average to translate an informally written social media text of eleven words from Chinese into English. On the other hand, it takes thirty seconds to normalize the same message, a six-fold increase in speed. After training a text normalization system to normalize social media texts, we can use an existing statistical machine translation (SMT) system trained on normal texts (non-social media texts) to carry out translation. So we argue that normalization followed by regular translation is a more practical approach. Thus, text normalization is important for social media text processing.

Most previous work on normalization of social media text focused on word substitution (Beaufort et al., 2010; Gouws et al., 2011; Han and Baldwin, 2011; Liu et al., 2012). However, we argue that some other normalization operations besides word substitution are also critical for subsequent natural language processing (NLP) applications, such as missing word recovery (e.g., zero pronouns) and punctuation correction.

In this paper, we propose a *novel* beam-search decoder for normalization of social media text for MT. Our decoder can effectively integrate different normalization operations together. In contrast to previous work, some of our normalization operations are specifically designed for MT, e.g., missing word recovery based on conditional random fields

Proceedings of NAACL-HLT 2013, pages 471–481,
Atlanta, Georgia, 9–14 June 2013.

(CRF) (Lafferty et al., 2001) and punctuation correction based on dynamic conditional random fields (DCRF) (Sutton et al., 2004).

To the best of our knowledge, our work is the first to perform missing word recovery and punctuation correction for normalization of social media text, and also the first to perform message-level normalization of Chinese social media text. We investigate the effects on translating social media text after addressing various characteristics of informal social media text through normalization. To show the applicability of our normalization approach for different languages, we experiment with two languages, Chinese and English. We achieved statistically significant improvements over two strong baselines: an improvement of 9.98%/7.35% in BLEU scores for normalization of Chinese/English social media text, and an improvement of 1.38%/1.35% in BLEU scores for translation of Chinese/English social media text. We created two corpora: a Chinese corpus containing 1,000 Weibo[1] messages with their normalizations and English translations; and another similar English corpus containing 2,000 SMS messages from the NUS SMS corpus (How and Kan, 2005). As far as we know, our corpora are the first publicly available Chinese/English corpora for normalization and translation of social media text[2].

2 Related Work

Zhu et al. (2007) performed text normalization of informally written email messages using CRF (Lafferty et al., 2001). Due to its importance, normalization of social media text has been extensively studied recently. Aw et al. (2005) proposed a noisy channel model consisting of different operations: substitution of non-standard acronyms, deletion of flavor words, and insertion of auxiliary verbs and subject pronouns. Choudhury et al. (2007) used hidden Markov model to perform word-level normalization. Kobus et al. (2008) combined MT and automatic speech recognition (ASR) to better normalize French SMS message. Cook and Stevenson (2009) used an unsupervised noisy channel model considering different word formation processes. Han and Baldwin (2011) normalized informal words using morphophonemic similarity. Pennell and Liu (2011) only dealt with SMS abbreviations. Xue et al. (2011) normalized social media texts incorporating orthographic, phonetic, contextual, and acronym factors. Liu et al. (2012) designed a system combining different human perspectives to perform word-level normalization. Oliva et al. (2013) normalized Spanish SMS messages using a normalization and a phonetic dictionary. For normalization of Chinese social media text, Xia et al. (2005) investigated informal phrase detection, and Li and Yarowsky (2008) mined informal-formal phrase pairs from Web corpora.

All the above work focused on normalizing words. In contrast, our work also performs other normalization operations such as missing word recovery and punctuation correction, to further improve machine translation. Previously, Aw et al. (2006) adopted phrase-based MT to perform SMS normalization, and required a relatively large number of manually normalized SMS messages. In contrast, our approach performs beam search at the sentence level, and does not require large training data.

We evaluate the success of social media text normalization in the context of machine translation, so research on machine translation of social media text is relevant to our work. However, there is not much comparative evaluation of social media text translation other than the Haitian Creole to English SMS translation task in the 2011 Workshop on Statistical Machine Translation (WMT 2011) (Callison-Burch et al., 2011). However, the setup of the WMT 2011 task is different from ours, in that the task provided parallel training data of SMS texts and their translations. As such, text normalization is not necessary in that task. For example, the best reported system in that task (Costa-jussà and Banchs, 2011) did not perform SMS message normalization.

In speech to speech translation (Paul, 2009; Nakov et al., 2009), the input texts contain wrongly transcribed words due to errors in automatic speech recognition, whereas social media texts contain abbreviations, new words, etc. Although the input texts in both cases deviate from normal texts, the exact deviations are different.

[1] A Chinese version of Twitter at `www.weibo.com`

[2] Available at www.comp.nus.edu.sg/∼nlp/corpora.html

Category	Freq.	Example
Punctuation	81	你好*[hi]* ～(你好 。 *[hi .]*);
Pronunciation	47	表*[watch]*(不要*[don't]*); 酱紫(这样子*[this]*);
New word	43	萌*[bud]*(可爱*[cute]*);
Interjection	27	好的*[ok]* 哦*[oh]*(好的*[ok]*);
Pronoun	23	想要*[want]*(我*[i]* 想要*[want]*);
Segmentation	14	表酱紫(不要*[don't]* 这样子*[this]*);
Pronunciation	288	*4*(*for*); *oredi*(*already*);
Abbreviation	98	*slp*(*sleep*); *whr*(*where*);
Prefix	74	*lect*(*lecture*); *doin*(*doing*);
Punctuation	69	*where r u*(*where r u ?*);
Interjection	68	*ok lor .*(*ok .*);
Quotation	24	*im sure*(*i 'm sure*); *dont go*(*don 't go*);
Be	24	*i coming*; *you free?*;
Tokenization	19	*ok.why ?*(*ok . why ?*);
Time	2	*end at 730*(*end at 7:30*); *1130 am*(*11:30 am*);

Table 1: Occurrence frequency of various informal characteristics in 200 Chinese/English social media texts.

3 Challenges in Normalization of Social Media Text

To better understand the informal characteristics of social media texts, we first analyzed a small sample of such texts in Chinese and English. We crawled 200 Chinese messages from Weibo. The informal characteristics of these messages are shown in the first half of Table 1. The manually normalized form is shown in round brackets, and the English gloss is shown in square brackets. Omitted, extraneous, and misused punctuation symbols occur frequently. On average, each Chinese message contains only less than one informal word, and many informal words are either new words or existing words with new meaning. The messages also contain redundant interjections like "哦*[oh]*". Pronouns are often omitted in Chinese messages, especially for "我*[I]*". Chinese informal words can be wrongly segmented due to lack of word segmentation training data containing informal words.

Similarly, 200 English SMS messages were randomly selected from the NUS SMS corpus (How and Kan, 2005). The informal characteristics of these messages are shown in the second half of Table 1. We found that our English messages contain more informal words than Chinese messages. English words are shortened in three ways: (1) using a shorter word form with similar pronunciation; (2) abbreviating a formal word; and (3) using only a prefix of a formal word. Other informal characteristics include: (1) informal punctuation conventions including omitted and misused punctuation; (2) redundant interjections; (3) quotation-related problems, e.g., omitted quotation marks; (4) "*be*" omission; (5) tokenization problems; and (6) informally written time expressions.

4 Methods

As can be seen in Section 3, social media texts of different languages exhibit different informal characteristics. For example, English messages have more informal words than Chinese messages, while punctuation problems are more prevalent for Chinese messages. Also, fixing different types of informal characteristics often depends on each other. For example, to be able to correct punctuation, it helps that the surrounding words are already correctly normalized. On the other hand, with punctuation already corrected, it will be easier to normalize the surrounding words.

In this section, we first present our punctuation correction method based on a DCRF model, and then present missing word recovery based on a CRF model. Next, we present a novel beam-search decoder for normalization of social media text, which can effectively integrate different normalization operations, including statistical and rule-based normalization. Finally, details of text normalization for Chinese and English are presented.

4.1 Punctuation Correction

In normalization of social media text, punctuation correction is also important besides word normalization, as the subsequent NLP applications are typically trained on formal texts with correct punctuation. We define punctuation correction as correcting punctuation in sentences which may have no or unreliable punctuation. The task performs three punctuation operations: insertion, deletion, and substitution.

To our knowledge, no previous work has been done on punctuation correction for normalization of social media text. In ASR, punctuation prediction only inserts punctuation symbols into ASR output that has no punctuation (Kim and Woodland, 2001; Huang and Zweig, 2002), but without punctuation deletion or substitution. Lu and Ng (2010) argued that punctuation prediction should be jointly per-

formed with sentence boundary detection, so they modeled punctuation prediction using a two-layer DCRF model (Sutton et al., 2004).

We also believe that punctuation correction is closely related to sentence boundary detection. Thus, we propose a two-layer DCRF model for punctuation correction. Layer 1 gives the actual punctuation tags *None*, *Comma*, *Period*, *Question-Mark*, and *Exclamatory-Mark*. Layer 2 gives the sentence boundary, including tags *Declarative-Begin*, *Declarative-In*, *Question-Begin*, *Question-In*, *Exclamatory-Begin*, and *Exclamatory-In*, indicating whether the current word is at the beginning of (or inside) a declarative, question, or exclamatory sentence.

We use word n-grams ($n = 1, 2, 3$) and punctuation symbols within 5 words before and after the current word as binary features in the DCRF model. As an example, Table 2 shows the tags and features for the word "*where*" in the message "*where| .|? i| can| not| see| you| !|!*", where the punctuation symbols after the vertical bars are the corrected symbols.

Tags	**Content**
Layer 1	*Question-Mark*
Layer 2	*Question-Begin*
Features	**Content**
unigram	<s>@-1 where@0 i@1 can@2 not@3 see@4 you@5
bigram	<s>+where@-1 where+i@0 i+can@1 can+not@2 not+see@3 see+you@4 you+</s>@5
trigram	<s>+where+i@-1 where+i+can@0 i+can+not@1 can+not+see@2 not+see+you@3 see+you+</s>@4
punctuation	.@0 !@5

Table 2: An example of tags and features used in punctuation correction.

Due to the lack of informal training texts with corrected punctuation, we train our punctuation correction model on formal texts with synthetically created punctuation errors. We randomly add, delete, and substitute punctuation symbols in formal texts with equal probabilities. Specifically, for $s \in \{, .?!\}$, $P(none|s) = P(,|s) = P(.|s) = P(?|s) = P(!|s) = 0.2$ denotes the probability of replacing a punctuation symbol s (replacing s by $none$ denotes deletion); and for a real word (not a punctuation symbol) w, $P(none|w) = P(,|w) = P(.|w) = P(?|w) = P(!|w) = 0.2$ denotes the probability of inserting a punctuation symbol after w (inserting $none$ after w denotes no insertion).

4.2 Missing Word Recovery

As shown in Section 3, some words are often omitted in social media texts, e.g., the pronoun "我*[I]*" in Chinese and *be* in English. To fix this problem, we propose a CRF model to recover such missing words. We explain the CRF model using *be* in English. The CRF model has five tags: *None*, *BE*, *IS*, *ARE*, and *AM*. In an input sentence, every token (including words, punctuation symbols, and a special start-of-sentence placeholder) will be assigned a tag, denoting the insertion of the form of *be* after the token. We use the same n-gram features as our punctuation correction model, but exclude the punctuation features. The model is trained on synthetically created training texts in which *be* has been randomly deleted with probability 0.5.

4.3 A Decoder for Text Normalization

When designing our text normalization system, we aim for a general framework that can be applied to text normalization across different languages with minimal effort. This is a challenging task, since social media texts in different languages exhibit different informal characteristics, as illustrated in Section 3. Motivated by the beam-search decoders for SMT (Koehn et al., 2007), ASR (Young et al., 2002), and grammatical error correction (Dahlmeier and Ng, 2012), we propose a novel beam-search decoder for normalization of social media text.

Given an input message, the normalization decoder searches for its best normalization, i.e., the best hypothesis, by iteratively performing two subtasks: (1) producing new sentence-level hypotheses from hypotheses in the current stack, carried out by *hypothesis producers*; and (2) evaluating the new hypotheses to retain good ones, carried out by *feature functions*. Each hypothesis is the result of applying successive normalization operations on the initial input message, where each normalization operation is carried out by one hypothesis producer that deals with one aspect of the informal characteristics of social media text. The hypotheses are grouped into stacks, where stack i stores all hypotheses obtained by applying i hypothesis producers on the input message. The beam-search algorithm is shown

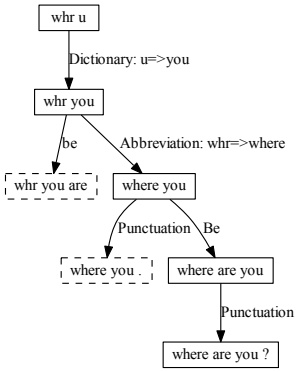

Figure 1: An example search tree when normalizing "*whr u*". The solid (dashed) boxes represent good (bad) hypotheses. The hypothesis producers are indicated on the edges.

in Algorithm 1, and Figure 1 shows an example search tree for an English message.

Algorithm 1 The beam-search decoder

INPUT: a raw message **M** whose length is **N**
RETURN: the best normalization for **M**

```
initialize hypothesisStacks[N+1] and hypothesisProducers;
add the initial hypothesis M to stack hypothesisStacks[0];
for i ← 0 to N-1 do
    for each hypo in hypothesisStacks[i] do
        for each producer in hypothesisProducers do
            for each newHypo produced by producer from hypo do
                add newHypo to hypothesisStacks[i+1];
                prune hypothesisStacks[i+1];
return the best hypothesis in hypothesisStacks[0...N];
```

We give the details of the hypothesis producers for Chinese and English social media texts in the next two subsections. A number of the hypothesis producers detect and deal with informal words w present in a hypothesis by relying on bigram counts of w in a large corpus of formal texts. Specifically, a word w in a hypothesis $\ldots w_{-1}ww_1\ldots$ is considered an informal word if both bigrams $w_{-1}w$ and ww_1 occur infrequently (≤ 5) in the formal corpus.

Given a hypothesis message h, the feature functions include a language model score (the normalized sentence probability of h), an informal word count penalty (the number of informal words detected in h), and count feature functions. Each count feature function gives the count of the modifications made by a hypothesis producer. The feature functions are used by the decoder to distinguish good hypotheses from bad ones. All feature functions are combined using a linear model to obtain the score for a hypothesis h:

$$score(h) = \sum_i \lambda_i f_i(h), \quad (1)$$

where f_i is the i-th feature function with weight λ_i. The weights of the feature functions are tuned using the pairwise ranking optimization algorithm (Hopkins and May, 2011) on the development set.

4.4 Text Normalization for Chinese

Taking into account the informal characteristics of Chinese social media text in Section 3, we design the following **hypothesis producers** for Chinese text normalization:

Dictionary: We have manually assembled a dictionary of 703 informal-formal word pairs from the Internet. The word pairs are used to produce new hypotheses. For example, given a hypothesis "神马*[magical horse]* 时候*[time]*", if the dictionary contains the word pair "(神马,什么*[what]*)", the Dictionary hypothesis producer generates a new hypothesis "什么*[what]* 时候*[time]*".

Punctuation: A punctuation correction model (Section 4.1) is adopted to correct punctuation in the current hypothesis, e.g., it may normalize "什么*[what]* 时候*[time]*" into "什么 时候？".

Pronunciation: We use Chinese Pinyin to model the pronunciation similarity of words. To accomplish this, we pair some Pinyin initials that sound similar into a group. The groups of paired Pinyin initials are *(c, ch)*, *(s, sh)*, and *(z, zh)*. For example, given the hypothesis "北京*[Beijing]* 筒子*[tube]* 来了*[come]*", the Pinyin of the informal word "筒子" is "*t ong z i*". The Pinyin of the formal word "同志*[comrade]*" is "*t ong zh i*". Since the similar sounding Pinyin initials *z* and *zh* are paired in a group, a new hypothesis "北京*[Beijing]* 同志*[comrade]* 来了*[come]*" can be produced.

In practice, this hypothesis producer can propose many spurious candidates w' for an informal word w. As such, after we replace w by w' in the hypothesis, we require that some 4-gram containing w' and its surrounding words in the hypothesis appears in a formal corpus. We call this filtering process *contextual filtering*.

Pronoun: With the method of Section 4.2, a CRF model is trained to recover the missing pronoun "我*[I]*".

Interjection: If a word w in a pre-defined list of frequent redundant interjections appears at the end of a sentence, we produce a new hypothesis by removing w, e.g., from "好的*[ok]* 哦*[oh]*" to "好的*[ok]*".

Resegmentation: This hypothesis producer fixes word segmentation problems. If an informal word is a concatenation of two constituent informal words w_1 and w_2 in our normalization dictionary, the informal word will be segmented into two words w_1 and w_2. As a result, the Dictionary hypothesis producer can subsequently normalize w_1 and w_2.

4.5 Text Normalization for English

Similar to Chinese text normalization, we also create the Dictionary, Punctuation, and Interjection hypothesis producers for English text normalization. We also add the following English-specific hypothesis producers:

Pronunciation: This hypothesis producer uses pronunciation similarity to find formal candidates for a given informal word. It considers a word as a sequence of letters and convert it into a sequence of phones using phrase-based SMT trained on the CMU pronouncing dictionary (Weide, 1998). Similar sounding phones are paired together in a group: *(ah, ao)*, *(ow, uw)*, and *(s, z)*. To illustrate, in the hypothesis "*wat is it*", the informal word "*wat*" maps to the phone sequence "*w ao t*". Since the formal word "*what*" maps to the phone sequence "*w ah t*" and the phones *ah* and *ao* are paired in a group, the new hypothesis "*what is it*" is generated.

Be: We train a CRF model to recover missing words *be*, as described in Section 4.2.

Retokenization: This hypothesis producer fixes tokenization problems. More precisely, given an informal word which is not a URL or email address and contains a period, it splits the informal word at the period. For example, "*how r u.where r u*" is normalized to "*how r u . where r u*".

Prefix: This hypothesis producer generates a formal word w' for an informal word w if w is a prefix of w'. To avoid spurious candidates, we only generate w' if $|w| \geq 3$ and $|w'| - |w| \leq 4$.

Quotation: If an informal word ends with a letter in *(m, s, t)* and if the word produced by inserting a quotation mark before the letter is a formal word, a new hypothesis with the quotation mark inserted is produced. This hypothesis producer thus generates "*i'm*" from "*im*", "*she's*" from "*shes*", "*isn't*" from "*isnt*", etc.

Abbreviation: Letters denoting the vowels in a formal word are often deleted to form an informal word. This hypothesis producer generates a formal word w' from an informal word w if w' can be obtained from w by adding missing vowels. To avoid spurious candidates, we only consider w where $|w| \geq 2$.

Time: If a number can be a potential time expression and appears after "*at*" or before "*am*" or "*pm*", a new hypothesis is produced by changing the number into a time expression, e.g., "*1130 am*" is normalized to "*11 : 30 am*".

Since the Pronunciation, Prefix, and Abbreviation hypothesis producers can propose spurious candidates for an informal word, we also use contextual filtering to further filter the candidates for these hypothesis producers.

5 Experiments

5.1 Evaluation Corpora

As previous work (Choudhury et al., 2007; Han and Baldwin, 2011; Liu et al., 2012) mostly focused on word normalization, no data is available with corrected punctuation and recovered missing words. We thus create the following two corpora (Table 3):

Chinese-English corpus We crawled 1,000 messages from Weibo which were first normalized into formal Chinese and then translated into formal English. The first half of the corpus serves as our development set to tune our text normalization decoder for Chinese, while the second half serves as the test set to evaluate text normalization for Chinese and Chinese-English MT.

English-Chinese corpus From the NUS English SMS corpus (How and Kan, 2005), we randomly selected 2,000 messages. The messages were first normalized into formal English and then translated into formal Chinese. Similar to the Chinese-English corpus, the first half of the corpus serves as our development set while the second half serves as the test set.

Corpus	# messages	# tokens (EN/CN/NCN)
CN2EN-dev	500	6.95K/5.45K/5.70K
CN2EN-test	500	7.14K/5.64K/5.82K
Corpus	**# messages**	**# tokens (EN/CN/NEN)**
EN2CN-dev	1,000	16.63K/18.14K/18.21K
EN2CN-test	1,000	16.14K/17.69K/17.76K

Table 3: Statistics of the corpora. *CN2EN-dev*/*CN2EN-test* is the development/test set in our Chinese-English experiments. *EN2CN-dev*/*EN2CN-test* is the development/test set in our English-Chinese experiments. *NEN*/*NCN* denotes manually normalized English/Chinese texts.

The formal corpus used (as described in Section 4) is the concatenation of two Chinese-English spoken parallel corpora: the IWSLT 2009 corpus (Paul, 2009) and another spoken text corpus collected at the Harbin Institute of Technology[3]. The language model used for Chinese (English) text normalization is the Chinese (English) side of the formal corpus and the LDC Chinese (English) Gigaword corpus.

To evaluate the effect of text normalization on MT, we build phrase-based MT systems using Moses (Koehn et al., 2007), with word alignments generated by GIZA++ (Och and Ney, 2003). The MT training data contains the above formal corpus and some LDC[4] parallel corpora (LDC2000T46, LDC2002E18, LDC2003E14, LDC2004E12, LDC2005T06, LDC2005T10, LDC2007T23, LDC2008T06, LDC2008T08, LDC2008T18, LDC2009T02, LDC2009T06, LDC2009T15, LDC2010T03). In total, 214M/192M English/Chinese tokens are used to train our MT systems. The language model of the Chinese-English (English-Chinese) MT system is the English (Chinese) side of the FBIS corpus (LDC2003E14) and the English (Chinese) Gigaword corpus. Our MT systems are tuned on the manually normalized messages of our development sets.

Following (Aw et al., 2006; Oliva et al., 2013), we use BLEU scores (Papineni et al., 2002) to evaluate text normalization. We also use BLEU scores to evaluate MT quality. We use the sign test to determine statistical significance, for both text normalization and translation.

[3] http://mitlab.hit.edu.cn/

[4] http://www.ldc.upenn.edu/Catalog/

5.2 Baselines

We compare our text normalization decoder against three baseline methods for performing text normalization. We then send the respective normalized texts to the same MT system to evaluate the effect of text normalization on MT.

The simplest baseline for text normalization is one that does no text normalization. The raw text (un-normalized) is simply passed on to the MT system for translation. We call this baseline ORIGINAL.

The second baseline, LATTICE, is to use a lattice to normalize text. For each input message, a lattice is generated in which each informal word is augmented with its formal candidates taken from the same normalization dictionary (downloaded from Internet) used in our text normalization decoder. The lattice is then decoded by the same language model used in our text normalization decoder to generate the normalized text (Stolcke, 2002). Another possible way of using lattice is to directly feed the lattice to the MT system (Eidelman et al., 2011), but since in this paper, we assume that the MT system can only translate plain text, we leave this as future work.

The third baseline, PBMT, is a competitive baseline that performs text normalization via phrase-based MT, as proposed in Aw et al. (2006). Moses (Koehn et al., 2007) is used to perform text normalization, by "translating" un-normalized text to normalized text. The training data used is the same development set used in our text normalization decoder. The normalized text is then sent to our MT system for translation. This method was also used in the SMS translation task of WMT 2011 by (Stymne, 2011).

In the tables showing experimental results, normalization and translation BLEU scores that are significantly higher than ($p < 0.01$) the LATTICE or PBMT baseline are **in bold** or underlined, respectively.

5.3 Chinese-English Experimental Results

The Chinese-English normalization and translation results are shown in Table 4. The first group of experiments is the three baselines, and the second group is an oracle experiment using manually normalized messages as the output of text normaliza-

System	BLEU scores (%) Normalization	MT
ORIGINAL baseline	61.01	9.06
LATTICE baseline	74.52	11.50
PBMT baseline	76.77	12.65
ORACLE	100.00	15.04
Dictionary	**77.80**	12.35
Punctuation	65.95	9.63
Pronunciation	61.30	9.13
Pronoun	61.11	9.01
Interjection	61.05	9.14
Resegmentation	60.98	9.03
Dictionary	**77.80**	12.35
+Punctuation	**84.69**	**13.37**
+Pronunciation	**84.69**	**13.40**
+Pronoun	**84.96**	**13.50**
+Interjection	**85.33**	**13.68**
+Resegmentation	**86.75**	**14.03**

Table 4: Chinese-English experimental results.

tion which indicates the theoretical upper bounds of perfect normalization. In the normalization experiments, the ORIGINAL baseline gets a BLEU score of 61.01%, and the LATTICE baseline greatly improves the ORIGINAL baseline by 13.51%, which shows that the dictionary collected from the Internet is highly effective in text normalization. The PBMT baseline further improves the BLEU score by 2.25%. In the corresponding MT experiments, as the normalization BLEU scores increase, the MT BLEU scores also increase.

The third group is the isolated experiments, i.e., each experiment only uses one hypothesis producer. As expected, the individual hypothesis producers alone do not work well except the Dictionary hypothesis producer. One interesting discovery is that the Dictionary hypothesis producer outperforms the LATTICE baseline, which shows that our normalization decoder can utilize the dictionary more effectively, probably because of the additional features used in our normalization decoder such as the informal word penalty. The Resegmentation hypothesis producer alone worsens the BLEU scores, since it can only split informal words, and is designed to work together with other hypothesis producers to normalize words.

The last group is the combined experiments. We add each hypothesis producer in the order of its normalization effectiveness in the isolated experiments. Adding the Punctuation hypothesis producer greatly improves the BLEU scores of both normalization and translation, which confirms the importance of punctuation correction. The Pronoun and Interjection hypothesis producers also contribute some improvements. Finally, Resegmentation significantly improves the normalization/translation BLEU scores by 1.42%/0.35%. Compared with the isolated experiments, the combined experiments show that our normalization decoder can effectively integrate different hypothesis producers to achieve better performance for both text normalization and translation.

Overall, in the Chinese text normalization experiments, our normalization decoder outperforms the best baseline PBMT by 9.98% in BLEU score. In the Chinese-English MT experiments, the normalized texts output by our normalization decoder lead to improved translation quality compared to normalization by the PBMT baseline, by 1.38% in BLEU score.

5.4 English-Chinese Experimental Results

The English-Chinese normalization and translation results are shown in Table 5, with the same experimental setup as in the Chinese-English experiments.

The text normalization BLEU score of the ORIGINAL baseline is much lower in English compared to Chinese, since the English texts contain more informal words. Again, the individual hypothesis producers alone do not work well, except the Dictionary hypothesis producer. The Retokenization hypothesis producer greatly improves the normalization/translation BLEU scores by 2.37%/0.86%. The Punctuation hypothesis producer helps less for English compared to Chinese, suggesting that our Chinese texts contain noisier punctuation.

Overall, we achieved similar improvements in English text normalization and English-Chinese translation, and the improvements in BLEU scores are 7.35% and 1.35% respectively.

5.5 Further Analysis

The effect of contextual filtering. To measure the effect of contextual filtering proposed in Section 4.4, we ran our normalization decoder without contextual filtering. We obtained BLEU scores of 65.05%/22.38% in the English-Chinese experiments, which were lower than 66.54%/22.81% ob-

System	BLEU scores (%) Normalization	MT
ORIGINAL baseline	37.38	13.63
LATTICE baseline	56.98	20.56
PBMT baseline	59.19	21.46
ORACLE	100.00	28.48
Dictionary	**59.90**	**20.84**
Retokenization	38.79	14.06
Prefix	38.68	13.90
Interjection	38.37	13.92
Quotation	38.04	13.65
Abbreviation	37.94	13.74
Time	37.65	13.66
Pronunciation	37.62	13.80
Punctuation	37.62	13.79
Be	37.47	13.59
Dictionary	**59.90**	**20.84**
+Retokenization	**62.27**	**21.70**
+Prefix	**63.22**	**21.88**
+Interjection	**64.85**	**22.30**
+Quotation	**65.24**	**22.31**
+Abbreviation	**65.35**	**22.34**
+Time	**65.59**	**22.38**
+Pronunciation	**65.64**	**22.38**
+Punctuation	**66.38**	**22.74**
+Be	**66.54**	**22.81**

Table 5: English-Chinese experimental results.

tained with contextual filtering. This shows the beneficial effect of contextual filtering.

Decoding speed. The decoding speed of our text normalization decoder was 0.2 seconds per message on our test sets, using a 2.27 GHz Intel Xeon CPU with 32 GB memory.

The effect of text normalization decoder on MT. We manually analyzed the effect of our text normalization decoder on MT. For example, given the un-normalized English test message "*yeah must sign up , im in lt25*", our English-Chinese MT system translated it into "对*[yeah]* 必须*[must]* 签署*[sign up]* , im 在*[in]* lt25" On the other hand, our normalization decoder normalized it into "*yeah must sign up , i 'm in lt25 .*" which was then translated into "对 必须 签署 , 我 在 lt25 。" by our MT system. This example shows that our text normalization decoder uses word normalization and punctuation correction to improve translation.

6 Conclusion

This paper presents a novel beam-search decoder for normalization of social media text. Our decoder for text normalization effectively integrates multiple normalization operations. In our experiments, we achieved statistically significant improvements over two strong baselines: an improvement of 9.98%/7.35% in BLEU scores for normalization of Chinese/English social media text, and an improvement of 1.38%/1.35% in BLEU scores for translation of Chinese/English social media text. Future work can investigate how to more tightly integrate our beam-search decoder for text normalization with a standard MT decoder, e.g., by using a lattice or an n-best list.

Acknowledgments

We thank all the anonymous reviewers for their comments which have helped us improve this paper. This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

AiTi Aw, Min Zhang, PohKhim Yeo, ZhenZhen Fan, and Jian Su. 2005. Input normalization for an English-to-Chinese SMS translation system. In *Proceedings of the Tenth MT Summit*.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of ACL-COLING*.

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédrick Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of ACL*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of WMT*.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*, 10(3-4):157–174.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*.

Marta R. Costa-jussà and Rafael E. Banchs. 2011. The BM-I2R Haitian-Créole-to-English translation system description for the WMT 2011 evaluation campaign. In *Proceedings of WMT*.

Daniel Dahlmeier and Hwee Tou Ng. 2012. A beam-search decoder for grammatical error correction. In *Proceedings of EMNLP-CoNLL*.

Vladimir Eidelman, Kristy Hollingshead, and Philip Resnik. 2011. Noisy SMS machine translation in low-density languages. In *Proceedings of WMT*.

Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef Van Genabith. 2011. # hardtoparse: POS tagging and parsing the twitterverse. In *Proceedings of the AAAI Workshop On Analyzing Microtext*.

Stephan Gouws, Dirk Hovy, and Donald Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of ACL-HLT*.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of EMNLP*.

Yijue How and Min-Yen Kan. 2005. Optimizing predictive text entry for short message service on mobile phones. In *Proceedings of Human Computer Interfaces International*.

Jing Huang and Geoffrey Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *Proceedings of ICSLP*.

Ji Hwan Kim and P. C. Woodland. 2001. The use of prosody in a combined system for punctuation generation and speech recognition. In *Proceedings of Eurospeech*.

Catherine Kobus, François Yvon, and Géraldine Damnati. 2008. Normalizing SMS: are two metaphors better than one? In *Proceedings of COLING*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL Demo and Poster Sessions*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Zhifei Li and David Yarowsky. 2008. Mining and modeling relations between formal and informal Chinese phrases from web corpora. In *Proceedings of EMNLP*.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of ACL-HLT*.

Fei Liu, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. In *Proceedings of ACL*.

Wei Lu and Hwee Tou Ng. 2010. Better punctuation prediction with dynamic conditional random fields. In *Proceedings of EMNLP*.

Preslav Nakov, Chang Liu, Wei Lu, and Hwee Tou Ng. 2009. The NUS statistical machine translation system for IWSLT 2009. In *Proceedings of the International Workshop on Spoken Language Translation*.

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

J. Oliva, J. I. Serrano, M. D. Del Castillo, and Á. Igesias. 2013. A SMS normalization system integrating multiple grammatical resources. *Natural Language Engineering*, 19(1):121–141.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Michael Paul. 2009. Overview of the IWSLT 2009 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*.

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

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of EMNLP*.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.

Sara Stymne. 2011. Spell checking techniques for replacement of unknown words and data cleaning for Haitian Creole SMS translation. In *Proceedings of WMT*.

Charles Sutton, Khashayar Rohanimanesh, and Andrew McCallum. 2004. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *Proceedings of the 21st International Conference on Machine Learning*.

Robert L. Weide. 1998. The CMU pronouncing dictionary. *URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.

Yunqing Xia, Kam-Fai Wong, and Wei Gao. 2005. NIL is not nothing: Recognition of Chinese network informal language expressions. In *4th SIGHAN Workshop on Chinese Language Processing*.

Zhenzhen Xue, Dawei Yin, and Brian D. Davison. 2011. Normalizing microtext. In *Proceedings of the AAAI Workshop on Analyzing Microtext*.

Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. 2002. The HTK book. *Cambridge University Engineering Department*.

Conghui Zhu, Jie Tang, Hang Li, Hwee Tou Ng, and Tie-Jun Zhao. 2007. A unified tagging approach to text normalization. In *Proceedings of ACL*.