A Transition-based Model for Joint Segmentation, POS-tagging and Normalization

Tao Qian[1,3], **Yue Zhang**[2], **Meishan Zhang**[2]*, **Yafeng Ren**[1] **and Donghong Ji**[1]
[1]Computer School, Wuhan University, Wuhan, China
[2]Singapore University of Technology and Design
[3]College of Computer Science and Technology, Hubei University of Science and Technology, XianNing, China
{taoqian, renyafeng, dhji}@whu.edu.cn
{yue_zhang, meishan_zhang}@sutd.edu.sg

Abstract

We propose a transition-based model for joint word segmentation, POS tagging and text normalization. Different from previous methods, the model can be trained on standard text corpora, overcoming the lack of annotated microblog corpora. To evaluate our model, we develop an annotated corpus based on microblogs. Experimental results show that our joint model can help improve the performance of word segmentation on microblogs, giving an error reduction in segmentation accuracy of 12.02%, compared to the traditional approach.

1 Introduction

Microblogs, such as Twitter, SMS and Weibo, has become an important research topic in NLP. Previous work has shown that off-the-shelf NLP tools can perform poorly on microblogs (Foster et al., 2011; Gimpel et al., 2011; Han and Baldwin, 2011). One of the major challenges for microblog processing is the issue of informal words. For example, "tmrw" has been frequently used in tweets for "tomorrow", causing OOV problems.

Text normalization has been introduced as a pre-processing step for microblog processing, which transforms informal words into their standard forms. Most work in the literature focuses on English microblog normalization, treating it as a noisy channel problem (Pennell and Liu, 2014; Cook and Stevenson, 2009; Yang and Eisenstein, 2013) or a translation problem (Aw et al., 2006; Contractor et al., 2010; Li and Liu, 2012; Zhang et al., 2014c), and training models based on words.

Lack of annotated corpora, text normalization is more challenging for Chinese. Unlike English, Chinese informal words are more difficult to mechanically normalize for two main reasons. First, Chinese does not have word delimiters. Second, Chinese informal words manifest diversity, such as abbreviations, neologisms, unconventional spellings and phonetic substitutions. Intuitively, there is mutual dependency between Chinese word segmentation and normalization, and therefore two tasks should be solved jointly.

Wang and Kan (2013) proposed a joint model to process word segmentation and informal word detection. However, text normalization was not included in the joint model. Kaji et al (2014) proposed a joint model for word segmentation, POS tagging and normalization for Japanese Microblogs, which was trained on a partially annotated microblog corpus. Their method requires special annotation for text normalization, which can be expensive.

In this paper, we propose a joint model for Chinese text normalization, word-segmentation and POS tagging, which can be trained using standard segmentation and POS tagging annotation, overcoming the lack of an annotated corpus on Chinese microblogs. Our model is based on Zhang and Clark (2010), with an extended set of transition actions to handle joint normalization. In our model, word segmentation and POS tagging are based on normalized text transformed from informal text. Assuming that the majority of informal words can be normalized into formal equivalents (Han et al., 2012; Li and Yarowsky, 2008), we seek standard forms of informal words from an automatically constructed normalization dictionary.

To evaluate our model, we developed an annotated corpus of microblog texts. Results show that our model achieves the best performances on three tasks compared with several baseline systems.

2 Text Normalization

Text normalization is a relatively new research topic. There are no precise definitions of a text

*corresponding author

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1837–1846, Lisbon, Portugal, 17-21 September 2015.

normalization task that are widely accepted by researchers. The task is generally divided into three categories: lexical-level, sentence-level and discourse-level normalization. In this paper we focus on lexical-level normalization, which aims to transform informal words into their standard forms.

Lexical normalization can be regarded as a spelling correction problem. However, researches on spelling correction focus on typographic and cognitive/orthographic errors (Kukich, 1992), while text normalization focuses on lexical variants, such as phonetic substitutions, abbreviation and paraphrases.

Unlike English, for which informal words are detected according to whether they are out of vocabulary, Chinese informal words manifest diversity. Wang et al. (2013) divided informal words into three types: phonetic substitutions, abbreviations and neologisms. Li and Yarowsky (2008) classified them into four types: homophone, abbreviation, transliteration and others. Due to variant characteristics, they normalise informal words by training a model per type, leading to increased system complexity.

Research reveals that most lexical variants have an unambiguous standard form (Han et al., 2012; Li and Yarowsky, 2008). The validity of this assumption is also empirically assessed on our corpus annotation in Section 6.1. Based on this assumption, we seek standard forms of informal words from a constructed normalization dictionary, avoiding diversity on informal words.

3 Joint Segmentation and Normalization

3.1 Transition-based Segmentation

We adapt the segmenter of Zhang and Clark (2007) as our baseline segmenter. Given an input sentence x, the baseline segmenter finds a segmentation by maximizing:

$$F(x) = \underset{y \epsilon Gen(x)}{\operatorname{argmax}} Score(y) \quad (1)$$

where $Gen(x)$ denotes the set of all possible segmentations for an input sentence.

Zhang and Clark (2007) proposed a graph-based scoring model, with features based on complete words and word sequences. We adapt their method slightly, under a transition-based framework (Zhang and Clark, 2011), which gives us a consistent way of defining all models in this paper.

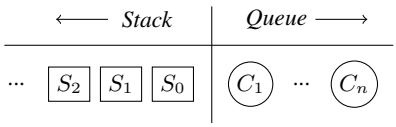

Figure 1: A state of transition-based model.

Here a transition model is defined as a quadruple $M = (C, T, W, C_t)$, where C is a state space, T is a set of transitions, each of which is a function: $C \rightarrow C$, W is an input sentence $c_1 ... \ c_n$, C_t is a set of terminal states. A model scores the output by scoring the corresponding transition sequence.

As shown in Figure 1, a state is a tuple $ST = (S, Q)$, where S contains partially segmented sequences, and $Q = (c_i, c_{i+1}, ..., c_n)$ is the sequence of input characters that have not been processed. When the character c_i is processing, the transition system would operate one of two actions that are defined as follows:

(1) APP(c_i), removing c_i from Q, and appending it to the last (partial) word in S.

(2) SEP(c_i), removing c_i from Q, making the last word in S as completed, and adding c_i as a new partial word.

Given the sentence "工作压力大啊！(How great work pressure is!)", the sequences of action "SEP(工), APP(作), SEP(压), APP(力), SEP(大), SEP(啊), SEP(!)" can be used to analyze its structure.

3.2 Joint Segmentation and Normalization

Our SN model extends the transition-based segmentation model. In addition to the actions APP and SEP, the transition system also contains a SEPS action, which substitutes an in formal word on the top of S if it exists in the normalization dictionary. Figure 2 gives a normalization transition process for the sentence "工作鸭梨大啊！(How great work pressure is!)". During processing the character "大(big)", the following actions can be applied.

(1) APP("大(big)"), appending "大(big)" to the last word "鸭梨(yālí, pear)" in the informal labeled sequence.

(2) SEP("大(big)"), making the last word "鸭梨(yālí, pear)" in the informal labeled sequence as a completed word, and adding "大(big)" as a new partial word.

(3) SEPS("大(big)", "压力(yālì, pressure)"), operating the action SEP("大(big)"), and using

Sentence: 工作鸭梨大啊! (How great work pressure is!)

State	Action	Stack	Queue	Dictionary
S_i	.....	Org: 工作 鸭梨 work pear Nor: 工作 work	大啊! big ah!	鸭梨- 压力 pear - pressure 孩纸- 孩子 child paper - child 围脖- 微博 neckerchief - microblog 盆友- 朋友 basin friend - friend .....
S_{i+1}	APP("大")	Org: 工作 鸭梨大 work pear big Nor: 工作 work	啊! (ah!)	
	SEP("大")	Org: 工作 鸭梨 大 work pear big Nor: 工作 work		
	SEPS("大", "压力")	Org: 工作 鸭梨 大 work pear big Nor: 工作 压力 work pressure		

Figure 2: Transition actions for joint segmentation and normalization.

Algorithm 1: Decoder

Input: sent, Dictionary // sent: informal sentence
Output:Best normalization sentence
1. *agenda* ←*NULL*
2. **for** *idx* ***in*** *[*0..LEN*(sent)]:*
3. **for** *state* ***in*** *agenda:*
4. *new* ←APP*(state, sent[idx])*
5. ADDITEM*(agenda, new)*
6. *new* ←SEP*(state, sent[idx])*
7. ADDITEM*(agenda, new)*
8. *norWords*←GETNWORD*(state.lastWord)*
9. **for** *word* ***in*** *norWords*
10. *new* ←SEPS*(state,sent[idx],word)*
12. ADDITEM*(agenda, new)*
13. *agenda* ←N-BEST*(agenda)*
14. **return** BEST*(agenda)*

the standard form "压力(yālì, pressure)" for the informal word " 鸭梨(yālí, pear)".

Given the sentence "工作鸭梨大啊！(How great work pressure is!)", the sequences of action "SEP(工), APP(作), SEP(鸭), APP(梨), SEPS(大, 压力), SEP(啊), SEP(!)" can be used to analyze its structure.

Lexical substitution is based on a normalization dictionary whose entries consist of <lexical variant, standard form> pairs. The output is a pair of labeled sequences, containing the informal labeled sequence and the corresponding formal labeled sequence. To rank the candidates, both labeled sequences can be scored. However, lacking annotated corpora on informal texts, we only use the score of formal labeled sequence in our model. The advantage is that we can train our model by using standard corpus only, overcoming the lack of annotated corpora on informal texts.

3.3 Training and Decoding

We apply the global training and beam-search decoding framework of Zhang and Clark (2011). An agenda is used by the decoder to keep the N-best states during the incremental process. Before decoding starts, the agenda is initialized with the initial state. When a character is processed, existing states are removed from the agenda and extended with all possible actions, and the N-best newly generated states are put back onto the agenda. After all states have been terminal, the highest-scored state from the agenda is taken as the output.

Algorithm 1 shows pseudocode for the decoder. ADDITEM adds a new item into the agenda, N-BEST returns the N highest-scored items from the agenda, and BEST returns the highest-scored item from the agenda. GETNWORD returns a possible standard form set of last word, seeking from normalization dictionary. APP appends a character to the last word in a state, SEP joins a character as the start of a new word in a state, SEPS operates SEP and replaces the last word by a possible standard form.

3.4 Features

In the experiments, we use the segmentation feature templates of Zhang and Clark (2011). These features are effective for segmentation on formal text. However, for text normalization, these features contain insufficient information. Our experiments show that by using Zhang and Clark's features, the F-Score on normalization is only 0.4207.

Prior work has shown that the language statistic information is important for text normalization (Wang et al., 2013; Li and Yarowsky, 2008; Kaji and Kitsuregawa, 2014). As a result, we extract language model features by using word-based language model learned from a large quantity of standard texts. In particular, 1-gram, 2-gram, 3-gram features are extracted. Every type of n-gram is divided into ten probability ranges. For example, if the probability of the word bigram: " 压力- 大" (high pressure) is in the 2_{nd} range, the feature is represented as "word-2-gram=2".

In our experiments, language models are trained on the Gigaword corpus[1] with SRILM tools[2]. To train a word-based language model, we segmented the corpus using our re-implementation of Zhang and Clark (2010). Results show that language model information not only improves the perfor-

[1]https://catalog.ldc.upenn.edu/LDC2003T05
[2]http://www.speech.sri.com/projects/srilm/

mance of text normalization, but also increases the performance of word-segmentation.

4 Extension for Joint Segmentation, Normalization and POS tagging

4.1 Joint Segmentation and POS Tagging

In order to reduce the error propagation of word segmentation, joint models have been applied to some NLP tasks, such as POS tagging (Zhang and Clark, 2010; Kruengkrai et al., 2009) and Parsing (Zhang et al., 2014a; Qian and Liu, 2012; Zhang et al., 2014b).

We take the joint word segmentation and POS tagging model of Zhang and Clark (2010) as the joint baseline. It extends from transition-based segmenter, adding POS arguments to the original actions. In Figure 1, when the current character c_i is processing, the transition system for ST would operate as follows :

(1) APP(c_i), removing c_i from Q, and appending it to the last (partial) word in S with the same POS tag, .

(2) SEP(c_i, pos), removing c_i from Q, making the last word in S as completed, and adding c_i as a new partial word with a POS tag "pos".

Given the sentence "工作压力大啊！(How great work pressure is!)", the sequences of action "SEP(工, NN), APP(作), SEP(压, NN), APP(力), SEP(大, VA), SEP(啊, SP), SEP(!, PU)" can be used to analyze its structure.

4.2 Joint Segmentation, Normalization and POS Tagging

Our joint model extends the model of Zhang and Clark (2010) by adding a SEPS action, which substitutes formal word for last word in S if exists in the dictionary. On the other hand, it can also be regarded as an extension of the joint segmentation and normalization model, adding POS arguments to the original actions.

Using the same example shown in Figure 2, the following three actions can be applied for the character " 大（big)":

(1) APP("大(big)"), appending "大(big)" to the last word "鸭梨(yālí, pear)" in the informal labeled sequence, which remain with the same POS tag "NN".

(2) SEP("大(big)", VA), making the last word "鸭 梨(yālí, pear)" in the informal labeled sequence as a completed word and adding "大(big)" as a new partial word with a POS tag "VA".

Text	Relation
海归也称海龟。(Overseas returnees is also referred to as turtles.)	(海归, 海龟) (overseas returnee, turtle)
一棵树，有点高，上面挂了好多人。恩，这棵树叫高数（高等数学）。(A tree, seemingly a little high, fails a lot of people. Well, this tree is called high number (advanced mathematics))	(高等数学, 高数) (advanced mathematics, high number)

Table 1: Relation patterns in microblogs.

(3) SEPS("大(big)", VA, "压 力(yālì, pressure)"), operating the action SEP("大(big)", VA), and using the standard form "压力(yālì, pressure)" for the informal word "鸭梨(yālí, pear)".

Given the sentence "工 作 鸭 梨 大 啊 ！ (How great work pressure is!)", the sequences of action "SEP(工, NN), APP(作), SEP(鸭, NN), APP(梨), SEPS(大, VA, 压 力), SEP(啊, SP), SEP(!, PU)" can be used to analyze its structure.

We use the same training and decoding framework as our joint segmentation, normalization and POS tagging model, as described in section 3.3.

5 Construction of Normalization Dictionary

Although large-scale normalization dictionaries are difficult to obtain, informal/formal relations could be extracted from large-scale web corpora (Li and Yarowsky, 2008), and informal words are mainly derived using fixed word-formation patterns. In this paper, we adopt two methods to construct a normalization dictionary.

The first method is to extract informal/formal pairs from large-scale text. In general, many informal and formal words co-occur in the same texts or similar contexts. We can find their relations with text patterns. As shown in Table 1, the first example follows the "formal也称informal" ("也称" means "is also referred to as") definition pattern, while the second example follows the pattern "informal(formal)". This gives us a reliable way to seed and bootstrap a list of informal/formal pairs.

We use a bootstrapping algorithm to extract informal/formal pairs from large-scale microblogs. First, a small set of example relations are collected manually. Second, using these relations as a seed set, we extract the text patterns, with which we identify more new relations from the data and aug-

informal也是formal的意思, formal也称informal, informal(formal), 为什么要把formal称为informal, formal其实叫informal,informal:...对formal的称谓, informal谐音自formal, “formal”缘何变“informal”, 用“informal”取代“formal”, informal就是formal的意思, informal有新的意义formal, formal的缩写是informal, 把formal说成informal, 网络...formal叫informal, 称...formal为informal, informal是formal的谐音, “formal”咋就成了“informal”, 当formal变成informal, informal可看作是formal的简称, 将formal写成informal, 网络...infromal的意思是formal.

Table 2: Examples of text patterns.

ment them into the seed set. Table 2 shows the initial text patterns extracted form the examples. The procedure iterates until it cannot identify new relations. There is much noise in the extracted informal/formal pairs. We re-rank them using a similarity-based classifier with weak supervision, with the positive pairs being inserted into dictionary.

The second method is to generate new informal/formal pairs using word-formation patterns extracted from informal/formal pairs. Although Chinese informal words manifest diversity, informal words are mainly derived using fixing word-formation methods, such as compounds, phonetic substitutions, abbreviations, acronym, reduplication. We can learn the pattern of informal word-formation from informal/formal pairs. For example, in informal/formal pair “ 妹纸(mèizhǐ, sister paper)/妹子(mèizǐ, sister)”, informal word “妹纸(mèizhǐ, sister paper)” is builded from formal word “妹子(mèizǐ, sister)” by the pattern “子→纸”. Using this pattern, we can generate many new informal/formal pairs, such as “汉纸(hànzhǐ, man paper)/汉子(hànzǐ, man)”, “男纸(nánzhǐ, man paper)/男子(nánzǐ, man)”, “孙纸(sūnzhǐ, grandson paper)/ 孙子(sūnzǐ, grandson)”, in which the formal words contain character “子”.

In the experiments, we constructed the normalization dictionary consisting of 32,787 informal/formal word pairs in total. The dictionary is used to tamper the formal training data for the joint segmentation and normalization systems with 25% of the formal words in the dictionary being replaced with their informal equivalents.

	Num	Ratio	Agree
Phonetic Substitutions	572	0.870	0.95
Abbreviation	69	0.105	0.97
Paraphrases	17	0.025	0.90
Total	658	1	0.95

Table 3: Frequency distribution and annotation agreement on various types of informal words.

6 Experiments

6.1 Microblog Corpus Annotation

To evaluate our model, we develop a microblog corpus. Our annotated corpus is collected from Sina Weibo[3], which is the largest microblogging platform in China. More than 1,000,000 Chinese posts are crawled using Sina Weibo API. Among these, 4,000 posts were randomly selected. We follow Wang et al. (2012) and apply rules to preprocess the corpus'URLs, emoticons, “@usernames” and Hashtags as pre-segmented words. As a result, we obtain 2,000 sentences as a source of the corpus.

Two human participants annotated the 2,000 sentences by using the tools we developed. The tools can simultaneously annotate word boundaries, POS and text normalization. We used the CTB scheme for word segmentation and POS tagging. We divided informal words into three types: Phonetic Substitutions, Abbreviation, Paraphrases. In total, we annotated 1,129 informal word-pairs in the 2,000 sentences, which contained 658 different informal words.

Table 3 shows the frequency distribution and annotation agreement over three types of informal words in corpus. The Cohen's kappa is 0.95 for informal words annotation, which shows that it is easy for humans to distinguish informal words, and validates our assumption that informal word generally has one formal word equivalent.

6.2 Settings and Measures

Our model is trained on the Chinese Treebank (CTB) 7[4], which is a large, word segmented, POS tagged and fully bracketed Chinese news corpus. The annotated microblog corpus is randomly divided into two parts: 1,000 sentences for development and 1,000 sentences for test.

The standard F-measure is used to measure the

[3]http://www.weibo.com/

[4]https://catalog.ldc.upenn.edu/LDC2010T07

	Development		Test	
	Seg-F	Nor-F	Seg-F	Nor-F
S;N	0.8859	0.3956	0.8885	0.4058
SN	0.8946	0.4053	0.8945	0.4207
S;N+lm	0.9101	0.5897	0.9132	0.6276
SN+lm	**0.9202**	**0.6009**	**0.9240**	**0.6392**

Table 4: Segmentation and normalization results. S;N denotes the pipeline model. SN denotes the joint model. lm denotes language model features.

	Segmentation		
models	N-R	I-R	ALL-R
S;N	0.8711	0.5100	0.8624
SN	0.8716	0.6653	0.8652
S;N+lm	0.9143	0.4229	0.9025
SN+lm	**0.9149**	**0.7752**	**0.9109**

Table 5: Formal and informal word accuracies on the development test. N-R denotes the recall rate of formal words, I-R denotes the recall rate of informal words, ALL-R denotes the recall rate of all words.

accuracies of word segmentation, POS tagging and text normalization, where the accuracy is *F = 2PR/(P+R)*. In addition, we use recall rates to evaluate the identification accuracies of formal, informal and all words. The recall rate of formal words *N-R* is defined as the percentage of gold standard output formal words that are correctly segmented, the recall rate of informal words *I-R* is defined as the percentage of gold-standard output informal words that are correctly segmented and the recall rate of all words *ALL-R* is defined as the percentage of gold standard output words that are correctly segmented.

6.3 Joint Segmentation and Normalization

Our development set is used to decide the beam size and the number of training iterations. The best performances on the development set are obtained when the beam size is set to 16 and the number of iterations is set to 32.

Comparison with pipeline We investigate the influence of the language model and analyze the result compared to the baseline. Table 4 shows the results on the development and test sets, where SN model is joint model and S;N is pipeline model. Our SN model performs better on segmentation than pipeline S;N model, demonstrating the effectiveness of normalization.

Table 5 shows the accuracies (i.e., recall rate) of formal and informal word identification on the development set. After normalization, the accuracy of informal word identification has a large improvement, and the accuracy of formal word identification also increases. This shows that formal words can be better recognized when informal words are identified correctly. It demonstrates that text normalization is effective for both informal words and formal words.

The effect of language model From Table 4, we observe that the performances increase when using language model features. Particularly, the normalization accuracy improves more significantly. It indicates that statistical language model knowledge play an important role on text normalization. Using language model features, our SN model improves more in the segmentation F-Score compared with the baseline system.

Furthermore, we also find that the language model features are helpful to identifying the formal words, as shown in Table 5. The identification accuracy of informal words increases on the SN model, while the accuracy decreases on the S;N model. Due to the relatively low frequency of informal words, they score lower on informal text by using the language model information, resulting in incorrect word segmentations. This illustrates that our joint model is more suitable for microblogs than the pipeline method.

6.4 Joint Segmentation, Normalization and POS tagging

We compare the following models on word segmentation, text normalization and POS tagging.

ST Our re-implementation of Zhang and Clark(2010). We investigate how the joint model contributes to improving accuracy of word segmentation and POS tagging in microblog domain.

S;N;T It is a pipe-line method for segmentation, normalization and POS tagging. The segmentation model does not use the features of POS. The normalization model uses segmentation information, but not features of POS. The POS tagging model does not need to segmentation.

SN;T It is another pipe-line method that first performs segmentation and normalization, then performs POS tagging. The SN model does not use the features of POS, and the POS tagging model does not need to segmentation.

SNT Our joint segmentation, normalization, and POS tagging model.

6.4.1 Results

Table 6 shows the final results on the test set. Previous work has shown that the systems trained on news data give poor accuracies of word segmentation and POS tagging in the microblog domain. As shown in Table 6,the F-Score of segmentation and POS tagging is 0.902 and 0.8163 respectively by using the Stanford segmenter and POS tagger.

Comparing ST and SNT, we find that text normalization can enhance word segmentation and POS tagging in the microblog. SNT achieved larger improvements over the baseline with language features, reducing segmentation errors by 12.02% and POS errors by 3.63%.

Another goal of the experiment is to illustrate whether the three tasks benefit from each other. Comparing SN;T to S;N;T shows that the performance increases by join segmentation and normalization. It indicates that segmentation and text normalization benefit from each other. On other hand, our SNT model yields better performance than SN;T. It indicates that POS features are effective for segmentation and text normalization, and hence three tasks benefit from each other.

The effect of the normalization dictionary The dictionary plays an important role in our model, which reduces the number of OOV words. Intuitively, the performance is higher when the coverage of dictionary is larger. In the experiments, the coverage of our dictionary on the development and tests are 45.8%,48.2% respectively.

To investigate the effect of the dictionary on our model, we manually construct ten dictionaries from our development data, with coverage between 10% and 100%. Figure 3 shows the F-score curves of test set on segmentation and POS-tagging for both SNT+lm and ST+lm model by different dictionaries. With the coverage of the dictionaries increasing from 10% to 100%, the F-score generally increases. When the coverage is greater than about 20%, the F-score for joint model is higher than for the baseline model.

6.4.2 Error Analysis

We found two major categories of errors. Abbreviation is sometimes incorrectly normalised, especially an informal word mapping to more than one formal word. For example, informal word “美偶” mapped to “美国偶像” (American idol), which consists of two words: “美国” (American) and “偶像” (idol). However, our model cannot normalise the word “美偶” in the experiment.

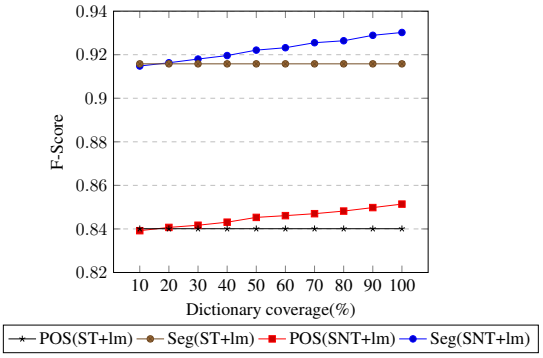

Figure 3: Results of SNT+lm and ST+lm based on different dictionaries for test set.

	Seg-F	POS-F	Nor-F
Stanford	**0.9058**	0.8163	
ST	0.8934	0.8263	
S;N;T	0.8885	0.8197	0.4058
SN;T	0.8945	0.8287	0.4207
SNT	0.8995	**0.8296**	**0.4391**
ST+lm	0.9162	0.8401	
S;N;T+lm	0.9132	0.8341	0.6276
SN;T+lm	0.9240	0.8439	0.6392
SNT+lm	**0.9261**	**0.8459**	**0.6413**

Table 6: Results on the test set. ST denotes the joint segmentation and POS tagging model. S;N;T denotes the pipeline model. SN denotes the joint segmentation and normalization model. SNT denotes the joint segmentatin. normalization and POS tagging model. lm denotes language model features. Seg-F denotes the F-Score of segmentation. POS-F denotes the F-Score of POS tagging. Nor-F denotes the F-Score of normalization.

Another type of error is phonetic substitutions of numbers, which are sometimes identified incorrectly. For example. “7456” is identified as a number in the experiments, but it means “气死我了” (I’m so angry). To settle this problem, it needs more context information.

6.5 Results of Lexical Normalization

It is interesting to explore how well the joint model can normalize informal words. We compare our results with two existing systems on text normalization based on our annotated microblog corpus.

(1) **WangDT** We re-implement Wang et al. (2013), which formalized the task as a classification problem and proposed rule-based and statistical features to model three plausible channels that explain the connection between formal and informal pairs. We use a single decision tree classifier

	P	R	F
SNT+lm	**0.9027**	0.4920	**0.6413**
WangDT	0.6214	**0.5543**	0.5859
LYTop1	0.6338	0.4920	0.5540

Table 7: Results of lexical normalization.

in the experiment.

(2) **LYTop1** Li and Yarowsky (2008) formalized the task as a ranking problem and proposed a conditional log-linear model to normalization. In the experiment, we select top 1 as the standard form of informal word.

We use the same division with 1000 sentences for training and 1000 for test. The training data is used for both the WangDT and LY. We re-segment the corpus using Stanford tools for the two baselines. WangDT uses CRF to detection informal words and LYTop1 uses the informal words detected using our joint model.

Although it is a little unfair for the two baselines compared with our joint model, which uses the external knowledge - normalization dictionary. The experiments can partly reflect some conclusions. Table 7 shows the results of normalization by different systems. The performance of our model is the best among the three systems. In particular, the precision in our SNT model improves upon the baselines significantly. The main reason is that our model is based on global features over whole sentences, while the two baselines based on local windows features.

7 Related Work

There has much work on text normalization. The task is generally treated as a noisy channel problem (Pennell and Liu, 2014; Cook and Stevenson, 2009; Yang and Eisenstein, 2013; Sonmez and Ozgur, 2014) or a translation problem (Aw et al., 2006; Contractor et al., 2010; Li and Liu, 2012; Zhang et al., 2014c). For English, most recent work (Han and Baldwin, 2011; Gouws et al., 2011; Han et al., 2012) uses two-step unsupervised approaches to first detect and then normalize informal words. They aim to produce and use informal/formal word lexicons and mappings.

In processing Chinese informal text, Wong and Xia (2008) address the problem of informal words in bulletin board system (BBS) chats by employing pattern matching. Xia et al. (2005) also use SVM-based classification to recognize Chinese informal sentences chats. Both methods have their advantages: the learning-based method does better on recall, while the pattern matching performs better on precision.

Li and Yarowsky (2008) tackle the problem of identifying informal/formal Chinese word pairs by generating candidates from Baidu search engine and ranking using a conditional log-linear model. Zhang et al. (2014c) analyze the phenomena of mixed text in Chinese microblogs, proposing a two-stage method to normalise mixed texts. However, their models employ pipelined words segmentation, resulting in reduced performance.

Wang and Kan (2013) propose a joint model to process word segmentation and informal word detection. However, text normalization is split to another task (Wang et al., 2013). Our joint model process word segmentation, POS tagging and normalization simultaneously. Kaji et al. (2014) propose a joint model for word segmentation, POS tagging and normalization for Japanese Microblogs. Their model is trained on a partially annotated microblog corpus. In contrast, our model can be trained on existing annotated corpora in standard text.

Researchers have recently developed various microblog corpora annotated with rich linguistic information. Gimpel et al. (2011) and Foster et al. (2011) annotate English microblog posts with POS tags. Han and Baldwin (2011) release a microblog corpus annotated with normalized words. Duan et al. (2012) develop a Chinese microblog corpus annotated with segmentation for SIGHAN bakeoff. Wang et al. (2013) release a Chinese microblog corpus for word segmentation and informal word detection. However, there are no microblog corpora annotated Chinese word segmentation, POS tags, and normalized sentences.

Our work is alse related to the work of word segmentation (Zhang and Clark, 2007; Zhang et al., 2013; Chen et al., 2015) and joint word segmentation and POS-tagging (Jiang et al., 2008; Zhang and Clark, 2010). A comprehensive survey is out of the scope of this paper, but interested readers can refer to Pei et al. (Pei et al., 2014) for a recent literature review of the fields.

To evaluate our model, we develop an annotated microblog corpus with word segmentation, POS tags, and normalization. Furthermore, we train our model by using a standard segmented and POS tagged corpus. We also present a comprehensive evaluation in terms of precision and recall on our

microblog test corpus. Such an evaluation has not been conducted in previous work due to the lack of annotated corpora for Chinese microblogs.

8 Conclusion

We proposed a joint model of word segmentation, POS tagging and normalization, in which the three tasks benefit from each other. The model is trained on standard corpora, hence there is no need to retrain it for new microblog corpora. The results demonstrated that the model can improve the performance of word segmentation and POS tagging with text normalization on microblogs, and our model can benefit from the language statistical information, which is not suitable to segment word and tag POS directly for microblogs because of the relatively low frequency of informal words.

In our model, lexical substitution is based on a normalization dictionary, which avoids the diversity of informal words, simplifying this problem for real world applications. The codes of the joint model and data set are published at the website: `https://github.com/qtxcm/JointModelNSP`.

Acknowledgments

We thank all reviewers for the insightful comments. This work is supported by the State Key Program of National Natural Science Foundation of China (No.61133012), the National Natural Science Foundation of China (No.61373108, 61373056, 61202193), the Key Program of Natural Science Foundation of Hubei, China(No.2012FFA088), the National Philosophy Social Science Major Bidding Project of China (No.11&ZD189) and the Singapore Ministration and Education (MOE) AcRF project T2MOE201301.

References

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL*, pages 33–40.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, and Xuanjing Huang. 2015. Gated recursive neural network for chinese word segmentation. In *Proceedings of the 53rd ACL*, pages 1744–1753, July.

Danish Contractor, Tanveer A Faruquie, and L Venkata Subramaniam. 2010. Unsupervised cleansing of noisy text. In *Proceedings of the 23rd COLING*, pages 189–196.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *Proceedings of the workshop on computational approaches to linguistic creativity*, pages 71–78.

Huiming Duan, Zhifang Sui, Ye Tian, and Wenjie Li. 2012. The cips-sighan clp 2012 chinese word segmentation on microblog corpora bakeoff. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, Tianjin, China*, pages 35–40.

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 *AAAI 2011 Workshop on Analyzing Microtext*, pages 20–25.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th ACL*, pages 42–47.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, pages 20–29.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th ACL*, pages 368–378.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 EMNLP*, pages 421–432.

Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lü. 2008. A cascaded linear model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of ACL-08: HLT*, pages 897–904.

Nobuhiro Kaji and Masaru Kitsuregawa. 2014. Accurate word segmentation and pos tagging for japanese microblogs: Corpus annotation and joint modeling with lexical normalization. In *Proceedings of the 2014 EMNLP)*, pages 99–109, October.

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint chinese word segmentation and pos tagging. In *Proceedings of the Joint Conference of the 47th ACL and the 4th AFNLP*, pages 513–521.

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

Chen Li and Yang Liu. 2012. Normalization of text messages using character-and phone-based machine translation approaches. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Zhifei Li and David Yarowsky. 2008. Mining and modeling relations between formal and informal chinese phrases from web corpora. In *Proceedings of the 2008 EMNLP*, pages 1031–1040.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for chinese word segmentation. In *Proceedings of the 52nd ACL*, pages 293–303, June.

Deana L Pennell and Yang Liu. 2014. Normalization of informal text. *Computer Speech & Language*, 28(1):256–277.

Xian Qian and Yang Liu. 2012. Joint chinese word segmentation, pos tagging and parsing. In *Proceedings of the 2012 EMNLP*, pages 501–511, July.

Cagil Sonmez and Arzucan Ozgur. 2014. A graph-based approach for contextual text normalization. In *Proceedings of the 2014EMNLP*, pages 313–324.

Aobo Wang and Min-Yen Kan. 2013. Mining informal language from chinese microtext: Joint word recognition and segmentation. In *ACL (1)*, pages 731–741.

Aobo Wang, Tao Chen, and Min-Yen Kan. 2012. Re-tweeting from a linguistic perspective. In *Proceedings of the second workshop on language in social media*, pages 46–55.

Aobo Wang, Min-Yen Kan, Daniel Andrade, Takashi Onishi, and Kai Ishikawa. 2013. Chinese informal word normalization: an experimental study. In *Proceedings of IJCNLP*, pages 127–135.

Kam-Fai Wong and Yunqing Xia. 2008. Normalization of chinese chat language. *Language Resources and Evaluation*, 42(2):219–242.

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 at IJCNLP*, volume 5.

Yi Yang and Jacob Eisenstein. 2013. A log-linear model for unsupervised text normalization. In *EMNLP*, pages 61–72.

Yue Zhang and Stephen Clark. 2007. Chinese segmentation with a word-based perceptron algorithm. In *Proceedings of the 45th ACL*, pages 840–847, June.

Yue Zhang and Stephen Clark. 2010. A fast decoder for joint word segmentation and pos-tagging using a single discriminative model. In *Proceedings of the 2010 EMNLP*, pages 843–852.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.

Longkai Zhang, Houfeng Wang, Xu Sun, and Mairgup Mansur. 2013. Exploring representations from unlabeled data with co-training for Chinese word segmentation. In *Proceedings of the 2013 EMNLP*, pages 311–321, October.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014a. Character-level chinese dependency parsing. In *Proceedings of the 52nd ACL*, pages 1326–1336, June.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014b. Type-supervised domain adaptation for joint segmentation and pos-tagging. In *Proceedings of the 14th EACL*, pages 588–597.

Qi Zhang, Huan Chen, and Xuanjing Huang. 2014c. Chinese-english mixed text normalization. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 433–442. ACM.