Chinese-English Mixed Text Normalization

Qi Zhang, Huan Chen, Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, P.R.China
{qz, 12210240054, xjhuang}@fudan.edu.cn

ABSTRACT

Along with the expansion of globalization, multilingualism has become a popular social phenomenon. More than one language may occur in the context of a single conversation. This phenomenon is also prevalent in China. A huge variety of informal Chinese texts contain English words, especially in emails, social media, and other user generated informal contents. Since most of the existing natural language processing algorithms were designed for processing monolingual information, mixed multilingual texts cannot be well analyzed by them. Hence, it is of critical importance to preprocess the mixed texts before applying other tasks. In this paper, we firstly analyze the phenomena of mixed usage of Chinese and English in Chinese microblogs. Then, we detail the proposed two-stage method for normalizing mixed texts. We propose to use a noisy channel approach to translate in-vocabulary words into Chinese. For better incorporating the historical information of users, we introduce a novel user aware neural network language model. For the out-of-vocabulary words (such as pronunciations, informal expressions and et al.), we propose to use a graph-based unsupervised method to categorize them. Experimental results on a manually annotated microblog dataset demonstrate the effectiveness of the proposed method. We also evaluate three natural language parsers with and without using the proposed method as the preprocessing step. From the results, we can see that the proposed method can significantly benefit other NLP tasks in processing mixed text.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - Information Search and Retrieval

General Terms

Algorithms, Experimentation.

Keywords

Words Normalization, Chinese-English Mixed Text, User Aware Neural Network Language Model

WSDM'14, February 24–28, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2556195.2556228.

1. INTRODUCTION

With the rapidly growing of Internet and the needs of globalization, exposure of individuals to multiple languages is becoming increasingly frequent. It promotes needs for people to acquire additional languages. Multilingual speakers have even outnumbered monolingual speakers [26]. Code-switching, which is the use of more than one language in the context of a single conversation, occurs frequently especially in informal texts. Due to the drastically increasing of social medias, the amount of user generated content (UGC) is extensively growing. Therefore, the mixed usages of more than one languages becomes a social phenomenon.

In Chinese, among all kinds of informal language phenomena, the mixed usage of Chinese and English is one the most frequent types. Through analyzing 210 million microblogs collected from Sina Weibo[1], which is one of the most popular website providing microblogging service in China, we find that over 14.8% microblogs contain at least one English word. Moreover, these English words include not only nouns but also adjectives, adverbs, and even verbs. For example, let us consider the following example:

帮我 ***book*** 一下会议室

(Please help me book a meeting room)

The speaker uses "*book*" instead of its Chinese translation " 预订" to express his meaning.

Since existing natural language processing techniques (e.g. POS tagging, chunking, parsing, opinion mining, etc.) were designed for processing monolingual text, multilingual mixed texts cannot be well processed by these methods. Moreover, due to lack of annotated corpus for informal texts, the effectiveness of most state-of-the-art supervised models are high impacted in processing informal content. We evaluated the performances of Stanford Parser and Berkeley Parser, which both of are widely used for various applications, drop to 66.4% and 67.9% respectively in processing the POS of English words in the Chinese-English mixed microblogs. It also demonstrates the great importance of normalizing the mixed texts. Zhao et al. [36] have also noticed this issue and proposed to use dynamic features under sequence labeling framework to achieve POS tagging problem for mixed texts. However, their work only focused on a specific task, POS tagging, and can not be directly adopted to process other tasks. Moreover, creating training data for mixed texts or investigating novel algorithms specifically for different NLP tasks are all time-consuming and sometimes difficult to accomplish. We argue that this kind of approaches can be easily generalized.

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

Several existing works have been proposed to achieve the task from the perspective of normalizing general informal Chinese texts. Li and Yarowsky [16] proposed to use bootstrapping model to identify candidate informal phrases and use conditional log-linear model based on rule-based intuitions and data co-occurrence phenomena to rank candidates. Wang and Ng [33] introduced a beam-search decoder based normalization method for missing word recovery, punctuation correction, manually assembled dictionary based word correction, and resegmentation. However, these methods are mainly based on the assumption of frequent occurrence. According to our statistic based on microblogs collected from real online service, the occurrence of English words in mixed texts also follow Zipf's law. Therefore, lots of English words' frequency in the mixed texts are low. These infrequent English words cannot be well translated or categorized by these methods.

Although most Chinese-English mixed microblogs only contain a few of English words, this is still a very challenging task due to the following facts: 1) there are enormous number and various types of English words. According to our statistic, more than 149K distinct words are included in 2.6 million mixed microblogs. 2) the linguistic and syntactic usages of English words may different from their original one. Due to that these mixed texts usually follow the grammar of one language, Part-of-speech tags of some English words may even be changed.

In this paper, we take a normalization centric view of processing Chinese-English mixed texts. We propose a novel two-stage method to achieve the task. For in-vocabulary English words, we propose to translate them into Chinese. Out-of-vocabulary words (including , pronunciations, informal expressions, etc.) are classified into different categories, such as person name, organization name, and so on. With these steps, mixed texts can be processed by existing NLP methods with little additional efforts. The normalized texts are much more easily for monolingual speakers to understand. To the best of our knowledge, this is the first work focused on normalizing Chinese-English mixed texts.

We propose to use noisy-channel approach with neural network language model to translate in-vocabulary words. To capture the historical information of users, we propose a novel user aware neural network language model. For training the word-level translation model, we constructed a parallel corpus based on subtitles of movie and TV series. To categorize words, we propose to use a graph-based unsupervised method with a novel initialization technique. For evaluating the proposed method, we also manually constructed a labeled corpus. Experimental results show that the proposed method achieves better performances than state-of-the-art methods. The main contributions of this paper are as follows:

- We formalize the English word normalization problem in Chinese-English mixed text. At the extent of the authors knowledge, it is the first work focused on this topic.
- To incorporate the historical information of users, we propose a novel user aware neural network language model.
- We manually labeled a number of microblogs extracted from real-world microblogs for evaluation.

The remaining parts are organized as follows: Section 2 gives some brief descriptions of related works. In section 3, we describe the phenomena and analysis of mixed texts. Section 4 describes the proposed normalization methods. Experimental results and analyses are given in Section 5. Section 6 concludes the paper.

2. RELATED WORK

The research of text normalization can be traced back to the task of converting numbers, dates, etc. into the standard dictionary words. Along with the rapid increasing of user generated content, text normalization task has received much more attentions in recent years [1, 13, 16, 2, 18, 14, 10, 6]. In this paper we classify the related works into three categories: lexical normalization, named entity normalization and informal text processing.

2.1 Lexical Normalization

Aw et al. [1] treated the lexical normalisation problem as a translation problem from the informal language to formal English They also studied the differences among SMS normalization, general text normalization, spelling check and text paraphrasing. Based on the investigated phenomena of SMS messages, they adapted a phrase-based method to achieve the task.

Kobus et al. [13] studied the problem of normalizing the orthography of French SMS messages. They proposed machine translation based method and nondeterministic phonemic transduction based method.

Han and Baldwin [8] proposed a supervised method to detect ill-formed words and used morphophonemic similarity to generate correction candidates. Then, all candidates were ranked based on a number of features generated from noisy context and similarity between ill-formed words and candidates.

Liu et al. [18] proposed to use a broad coverage lexical normalization method consisting three components. They assumed that a set of letter transformation patterns were used by humans to decipher the nonstandard tokens and integrated three human perspectives, including enhanced letter transformation, visual priming, and string/phonetic similarity

Han et al. [9] introduced a dictionary based method and an automatically normalisation dictionary construction method. They assumed that lexical variants and their standard forms occur in similar contexts.

Derczynski et al. [6] also proposed to use dictionary based method to achieve the task. They created a set of mappings from OOV words to their IV equivalents, using slang dictionaries and manual examination of the training data.

Wang and Ng [33] focused on the problem of missing word recovery, punctuation correction, manually assembled dictionary based word correction, and resegmentation. They introduced a beam-search decoder based normalization method to do it.

Although these methods achieved significant improvement on processing SMS and UGCs, they only focused on the monolingual text. Hence, Chinese-English mixed text can not be directly processed by these methods.

2.2 Named Entity Normalization

Normalizing named entity abbreviations to their standard forms is also an important preprocessing task for UGCs. This task has also attracted lots of attentions [3, 23, 17, 16, 34, 35].

Chang and Teng [3] introduced an HMM-based single character recovery model to extract character level abbreviation pairs for textual corpus. Okazaki et al. [23] also used discriminative approach for this task. They formalized the abbreviation recognition task as a binary classification problem and used Support Vector Machines to model it.

Yang et al. [35] treated the abbreviation generation problem as a labeling task and used Conditional Random Fields (CRFs) to do it. They also proposed to rerank candidates by a length model and web information.

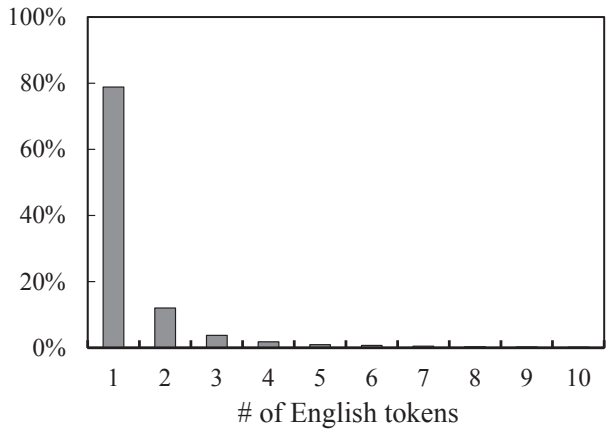

Figure 1: Distribution of number of English tokens per microblog.

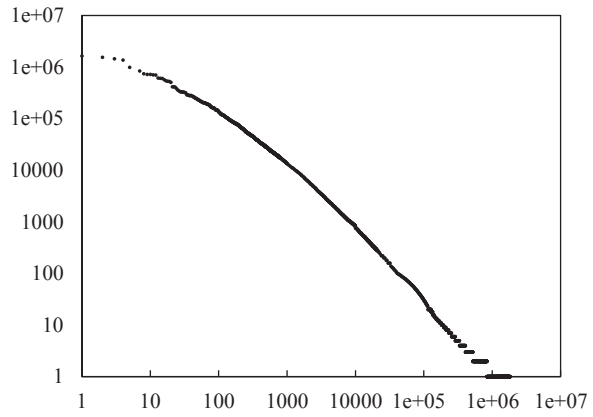

Figure 2: English token frequency in Chinese-English mixed microblogs.

Xie et al. [34] proposed to extract Chinese abbreviations and their corresponding definitions based on anchor texts. They constructed a weighted bipartite graph from anchor texts and applied co-frequency based measures to quantify the relatedness between two anchor texts.

Li and Yarowsky [17] proposed an unsupervised method extracting the relation between a full-form phrase and its abbreviation from monolingual corpora. They used data co-occurrence intuition to identify relations between abbreviation and full names. They also improved a statistical machine translation by incorporating the extracted relations into the baseline translation system.

Based on the data co-occurrence phenomena, Li and Yarowsky [16] also introduced a bootstrapping procedure to identify formal-informal relations in web corpora. They used search engine to extract contextual instances of the given informal phrase, and ranked the candidate relation pairs using conditional log-linear model.

2.3 Informal Text Processing

Despite the normalization based methods, a number of works have been proposed to directly process the informal texts [7, 19, 20, 30, 27, 36].

Freitag [7] studied the problem of performing information extraction from informal text. They showed the strategies of creating a term-space representation and exploiting typographic information in the form of token features. Minkov et al. [19] also introduced methods for extracting named entitles from informal texts and showed that informal text had different characteristics from formal text.

Mullen and Malouf [20] described statistical tests on a dataset of political discussion group postings. They concluded that traditional text classification methods would be inadequate to the task of sentiment analysis in this domain.

Thelwall et al. [30] focused on the task of detecting sentiment strength from informal texts. Through experimental results, they demonstrated incorporating decoding nonstandard spellings seemed to be one of factors of relative improvements given by their method.

Ritter et al. [27] experimental demonstrated that existing methods for POS tagging, Chunking and Named Entity Recognition performed quite poorly for processing Tweets. They presented a distantly supervised approach based LabeledLDA for Named Entity Classification problem on tweets.

Zhao et al. [36] proposed to use dynamic features under sequence labeling framework to process POS tagging problem for Chinese-English mixed texts. They extracted features from both local and non-local information and taked advantage of the characteristics of the mixed texts.

Previous works have been made to study the problem from various aspects. However, most works focused on specific tasks. Different with them, in this paper, we take a normalization centric view of processing Chinese-English mixed texts.

3. DATA ANALYSIS

For better understanding the phenomena of mixed usage of Chinese and English, in this section, we examine the dataset which contains about 210 millions microblogs crawled from Sina Weibo. We firstly describe the analyzing results from raw dataset. Then, we introduce the results acquired from manually categorized English words in these mixed texts. Since Chinese phonetic system and some informal usages are also represented by English alphabet letters, for clarification, we use *English token* to represent a sequence of English alphabet letters without any blank in between.

Firstly, the microblogs which contain at least one English token are extracted from the dataset. We observe that more than 14.8% microblogs are Chinese-English mixed texts. Figure 1 shows the distribution of number of English tokens per microblog in the Chinese-English mixed microblogs. We can observe that more than 94.6% microblogs contain less than 3 English tokens. About 78.8% mixed texts contain only one English token. It means that most English tokens are surrounding by Chinese characters. Hence, the linguistic usages of these English tokens may be different from their original ones. The part-of-speech tags of some tokens are even changed.

From the perspective of English token, we also look at the frequency of each token. Figure 2 shows a plot of English tokens frequency . The plot is in log-log coordinates. X-axile is the rank of a token in the frequency table. Y-axile is the total number of the token's occurrences. From the figure, we can observe that the frequency of English tokens also follow Zipf's law. It means that lots of tokens occur infrequently.

Table 1: Categories of English tokens in Chinese-English mixed texts. (English translations are in the brackets.)

Category	Percent	Example
Vocabulary word[a]	68.3%	别忘记明天的*meeting* (Please don't forget tomorrow's meeting).
Abbreviation	12.3%	*BBC*拍摄的《美丽中国》(Wild China produced by BBC).
Pronunciation	4.0%	发了几条*weibo* (Update several microblogs).
Slang	7.8%	*Orz* (A posture emoticon representing a kneeling person).
Other	7.6%	User ID, Chinese word followed by *ing*, misspelling, and so on.

[a]The vocabulary is constructed based on the parallel corpus, which will be described in Section 4.1.

For investigating the types of these English tokens, we randomly selected 2,000 microblogs which contain at lest one English token and manually labeled categories of English tokens. The five categories we use to classify queries are listed in Table 1. Table 1 also shows examples and percentages of each category. From the table, we can observe that vocabulary words and abbreviations take part in 68.3% and 12.3% among all English tokens respectively. It means that the tokens which can be translated take great part in all mixed texts. Among all the five categories, tokens belonging to "Slang" and "Other" categories are most difficult to normalize. For example, "Orz" is originated from Japan and can be used to express various meanings in different context.

4. THE PROPOSED METHOD

In this section, we describe the proposed normalization method for English tokens in Chinese-English mixed texts. For the tokens which are vocabulary words, we propose to use noisy channel model with word embeddings to translate them. For the out-of-vocabulary words and other types of English tokens, a graph-based unsupervised method is introduced to categorize them. The following sections will describe the proposed methods.

4.1 Word Translation

Let t represents the given mixed text. It contains a sequence of words $w_1w_2....w_n$. Each word w_i is either a Chinese word or an English word. For the mixed text t, the word translation method try to produce the normalization candidate $\hat{c}$ under the noisy channel model, which contain two components:

- A *language model* assigns a probability $p(c)$ for any sentence $c = w_1w_2....w_n$ in Chinese.
- A *translation model* assigns a conditional probability $p(c|t)$ to any Chinese/Mixed-Text pair of sentences.

Given these two components of the model, following the general noisy-channel approach, the output of the translation model on a Chinese-English mixed sentence t is:

$$\hat{c} = \arg\max_{c \in C} p(c) \times p(c|t), \tag{1}$$

where C is the set of all sentences in Chinese.

For language model, motivated by recent great success achieved by neural language models [11], we also incorporate it in this work. To better capture the historical information of users, we propose a novel user-aware neural language model. The proposed model learns to discriminate the next word given a short word sequence (local context) and sentences the user wrote recently (user historical information). As shown in Figure 3, two scoring components are defined for the final score of a (word sequence, user historical information) pair. The scoring components are computed by two neural networks.

Following the framework proposed by Collobert and Weston [4], given a word sequence c and user historical information u, our goal is to discriminate the correct last word in c for other random words. $s(c, u)$ represents the scoring function modeled by neural networks. c^w represents word sequence c with the last word replaced by word w. Hence, the objective is that the margin between $s(c, u)$ and $s(c^w, u)$ is larger than 1, for any other word w in the vocabulary. The object function is to minimize the ranking loss for for each (c, u) in the training corpus:

$$L_{c,u} = \sum_{w \in V} \max(0, 1 - s(c, u) + s(c^w, u)) \tag{2}$$

Firstly, the word sequence $c = w_1w_2....w_n$ is represented by an ordered list of vectors $x = (x_1, x_2, ..., x_n)$ where x_i is the embedding of word i in the sequence. x_i is a column in the embedding matrix $E \in \mathbb{R}^{m \times |V|}$, in which $|V|$ denotes the vocabulary size. The embedding matrix E will be learned and updated during the training procedure. $score_l$ is modeled by a neural network with one hidden layer:

$$a_1 = f(W_1[x_1; x_2; ...; x_n] + b_1) \tag{3}$$

$$score_l = W_2a_1 + b_2, \tag{4}$$

where f is an element-wise activation function such as $tanh$; $a_1 \in \mathbb{R}^{h \times 1}$ is the activation of the hidden layer with h hidden nodes; $W_1 \in \mathbb{R}^{h \times (mn)}$ is the first layer weights of the neural network; $W_2 \in \mathbb{R}^{1 \times h}$ is the second layer weights; b_1, b_2 are the biases of each layer.

Following the work done by Huang et al.[11], for representing user historical information, we also use the weighted average of all embeddings of words belonging to the user historical information. u denotes the user historical information vector and is calculated as follows:

$$u = \frac{\sum_{i=1}^{m} f(w_i^u) x_i^u}{\sum_{i=1}^{m} w(w_i^u)}, \tag{5}$$

where $w_1^u, w_2^u, ...w_m^u$ represents the words of user historical information; x_i^u denotes the embedding of w_i^u; $f(\cdot)$ captures the importance of the given word w_i. In this paper idf-weighting is used as the weighting function.

We also use a neural network with one hidden layer to compute the user historical information score, $score_u$ as follows:

$$a_1^u = f(W_1^u[u; x_n] + b_1^u) \tag{6}$$

$$score_u = W_2^u a_1^u + b_2^u, \tag{7}$$

where $[u; x_n]$ is the concatenation of the weighted average user historical information vector and the vector of the last word in t; $a_1^u \in \mathbb{R}^{h^u \times 1}$ is the activation of the hidden layer with h^u hidden

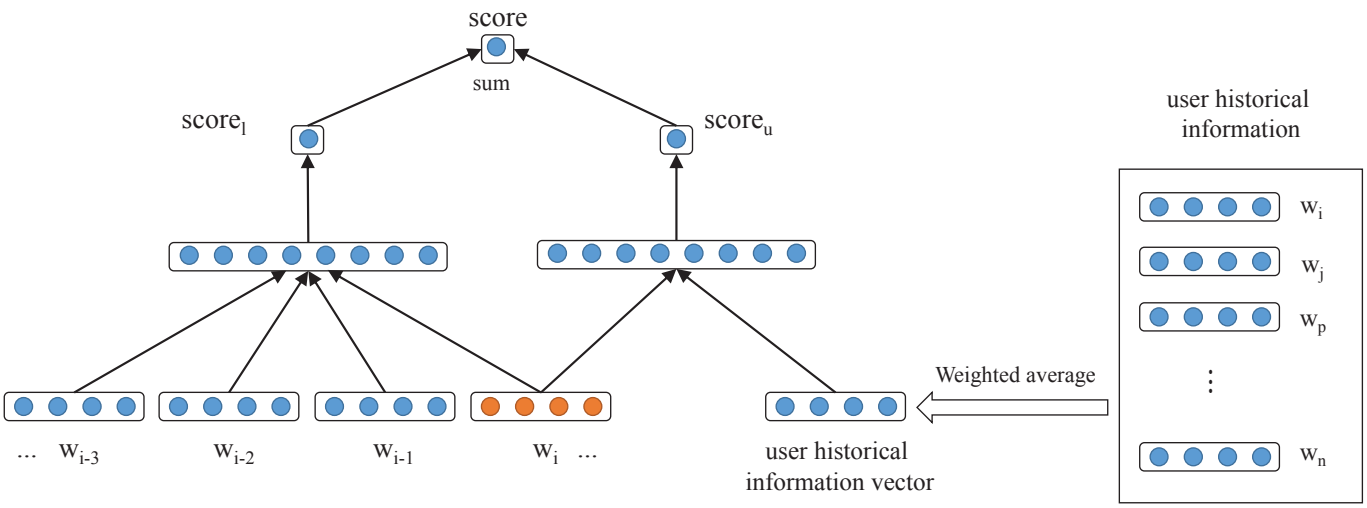

Figure 3: Overview structure of the proposed user-aware neural language model.

nodes; $W_1^u \in \mathbb{R}^{h^u \times (2m)}$ is the first layer weights of the neural network; $W_2^u \in \mathbb{R}^{1 \times h^u}$ is the second layer weights; b_1^u, b_2^u are the biases of each layer.

The final score is the sum of score of local context and user historical information:

$$score = score_l + score_u \tag{8}$$

For training the parameters: weights of the neural network and the embedding matrix E, we also follow the corrupt example sampling method [4], and sample the gradient of the objective by randomly choosing a word from the dictionary as a corrupt example for each sequence-context pair, (t, u). These weights are updated via back-propagation.

For the translation model $p(c|t)$, we make the following independence assumptions:

$$p(c|t) = \prod_{t_i \in Eng} p(c_i|t_i), \tag{9}$$

where $p(c|t)$ is the probability of generating Chinese word c from English word t and can be estimated by IBM Model 1 with parallel corpus.

Based on Eq.(8) and Eq.(9), the translation model Eq.(1) on a new Chinese-English mixed sentence t can be reformulated as:

$$\begin{aligned} \hat{c} &= \arg\max_{c \in C} p(c) \times p(c|t) \\ &\propto \arg\max_{c \in C} \prod_{i, t_i \in Eng} score(t_1 t_2 .. t_i, u) p(c_i|t_i) \end{aligned} \tag{10}$$

According to the statistic of the crawled corpus, we observe that more than 94.6% percent Chinese-English mixed texts contain less than 3 English words. Hence, in this work, the decoding problem can be efficiently solved.

4.2 Word Categorization

As described in Section 2, except in-vocabulary words, there are also a number of English tokens used as product name, informal expressions, and so on. For these tokens, we propose to use label propagation (LP) [37] to classify them into different categories. In this work, we try to classify English tokens into the following five categories: person name, product name, organization name, slang, and loanword.

From analyzing the dataset, we observe that (I) words belongs to the same categories tend to have similar contexts; (II) words and their corresponding category description words tend to have frequent co-occurrence relations. Previous researches also show that words with high context similarity tend to have similar semantic meanings [9]. Based on the observation, we propose to use context similarities to measure the edge weight of the graph. LP transfers labels from labeled data to unlabeled data through weighted graph. Based on the observation (II), we propose a novel label initialization method.

4.2.1 *Graph Construction*

We construct an undirected graph $G = (V, E)$ to represent the relations between English tokens. $V = \{v_1, ..., v_n\}$ denotes all the vertices in the graph. Vertices represent English tokens needed to be categorized. $E = \{e_{ij}, 1 \leq i, j \leq n\}$ represents the similarities between tokens. e_{ij} represents the similarity between vertex v_i and v_j. In order to reduce the computation cost of iteratively propagating, we also exclude edges whose value is less than threshold θ to prune the edges.

For calculating the similarities between tokens, we first extract all the context words in a predefined window (the windows size used in this work is 4) for each token. Context words are treated independently for each other. We use vector space model to construct context vector $\vec{f}$, of which each dimension is calculated by $tf \cdot idf$. Cosine measure is used to calculate the similarity between tokens:

$$e_{ij} = \cos(\vec{f_i}, \vec{f_j}) = \frac{\vec{f_i} \cdot \vec{f_j}}{\| \vec{f_i} \| \| \vec{f_j} \|} \tag{11}$$

4.2.2 *Label Propagation*

Label propagation method transfers the labels from labeled data to unlabeled data based on the constructed weighted graph. It has been successfully used for many tasks [21, 15, 32, 5, 29]. In this work, we also adopt LP to obtain the categories of English tokens.

From the observation (II), we know that words tend to have frequent co-occurrence with their category description words. For example, more than 1.98 billions documents can be retrieved using the query "iphone product" through Google. While, only 25 millions documents are returned using the query "iphone informal

expressions". Based on the observation, we firstly construct a number of description words for each category. cdw_{ij} represents jth description word of the ith category. Similar as SO-PMI-IR [31], we propose to use the following equation to measure the possibility of token v_i in the category z_j:

$$\text{SO-P}(v_i, z_j) = \max_{k=1}^{m} \frac{\text{hits}(v_i \text{ NEAR } cdw_{jk})}{\text{hits}(v_i) \cdot \text{hits}(cdw_{jk})}, \quad (12)$$

where hits($query$) is the number of hits given the query $query$ and is returned by Bing; "NEAR" is used to restrict the distance between search phrases[2].

The label distribution of each vertex is initialized as follows:

$$q_i^0(z_j) = \begin{cases} 1 & if\ v_i \in V_s\ and\ v_i \in V_{s_j} \\ 0 & if\ v_i \in V_s\ and\ v_i \notin V_{s_j} \\ \dfrac{\text{SO-P}(v_i, z_j)}{\sum_{k=1}^{m} \text{SO-P}(v_i, z_k)} & otherwise \end{cases}, \quad (13)$$

where $q_i^k(i = 1...|V|)$ is the category distribution for vertex v_i after k propagation; $q_i^k(z_j)$ represents the weight of a category z_j in q_i^k ; V_{s_j} is the set of seed words category for category z_j; $V_s = V_{s_1} \cup ... \cup V_{s_m}$ is the set of seed words of all categories.

With the initialization weights, label propagation method is used to iteratively update q_i^k through weighted edges. In each iteration, the probability propagation is also under the condition that edges with higher similarities allow easier propagation. Category distributions for each vertex is updated as follows:

$$q_i^k(z_j) = \begin{cases} q_i^0(z_j) & if\ v_i \in V_s \\ \dfrac{\sum_{v_j \in N(v_i)} e_{ij} \cdot q_j^{k-1}(z_j)}{\sum_{v_j \in N(v_i)} e_{ij}} & otherwise \end{cases}, \quad (14)$$

where $N(v_i)$ is the set of vertices linking to v_i.

5. EXPERIMENTS

In this section, we describe the experimental evaluations of the proposed method. Firstly, we describe the collections used for evaluation and experimental setups. Secondly, the performances of word translation and categorization are presented respectively. Finally, we evaluate performances of three Chinese parsers with and without the proposed method as preprocessing step.

5.1 Collection

As described in Section 2, for analyzing the phenomena of Chinese-English mixed text, we collected 210 millions microblogs from Sina Weibo. For evaluating the effectiveness of the proposed methods, we used a subset of them as testing data. We randomly selected 1,000 microblogs from all Chinese-English mixed ones and manually labeled the translation or categories of all English tokens in these texts. The testing data contains 1,195 English tokens in total.

Three annotators were involved in the labeling task. Since most mixed texts contain only a small number of English tokens, the ambiguity problem is not serious. Annotators were firstly asked to provide translations for all tokens. To evaluate the quality of corpus, we validate the agreements of human annotations using Cohen's kappa coefficient. The average κ among all annotators is 0.646. It indicates that the annotations of the corpus are reliable. Since some words have multiple translations, one of the annotator made the final decision to decide which translations should be included as the golden standards. For the word categorization, annotators were also asked to label categories for every words. If a category were labeled by more than two annotators for a word, the category is selected as the standard label of the word.

[2]We use the advanced keywords of Bing "near:10" to implement the NEAR constraint.

Table 2: Word translation results of different methods.

Methods	Accuracy (%)		
	Top-1	Top-5	Top-10
D-LM †	28.9	47.6	51.9
D-NLM	29.5	48.9	52.0
D-NLM+U	31.2	50.9	52.6
PC-LM †	60.3	80.7	84.0
PC-NLM	61.4	83.8	88.1
PC-NLM+U	**64.6**	**86.2**	**91.5**
Li and Yarowsky[16]	21.2	33.6	37.5
Han et al.[9]	19.6	27.2	31.3

† The in-vocabulary words based on online dictionary and parallel corpus take part in 66.3% and 67.6% respectively among all English tokens.

5.2 Experiment Configurations

For training the word translation model described in Section 3.1, we also collected 24,853 subtitles of movies and TV series from Shooter [3]. All these subtitles contain both Chinese and English text in single or separated files. Using these subtitles, we construct a parallel corpus, which contains more than 18.5 millions sentence pairs. For training the neural network language model, we randomly sampled 10 millions microblogs, due to the computation limited. We implement the proposed method based on the code of Huang et al. [11][4].

FudanNLP [25] is used for Chinese word segmentation. For training the word translation model, we use Giza++ toolkit [22] with the parallel corpus we constructed. For constructing the similarity graph, we implemented it using Hadoop 1.2.0 to handle massive computation. We also incorporate the proposed method as the preprocess step for three parsers: Stanford Parser 3.2 [28], Berkeley Parser 1.7 [12], and FudanNLP 1.57 [25].

For evaluating word translation, we adopt word-level n-best accuracy to evaluate the proposed methods. For each English token, the output is considered as correct if any of the corresponding golden standard words is among the top-n returned results. Evaluation metrics used for word categorization throughout the experiments include: Precision, Recall, and F1-score.

5.3 Word Translation Results

For translation model, we compare the proposed parallel corpus based method with dictionary[5] based method.

- "D" represents dictionary based method, where all the translations given by the dictionary are selected as candidates.
- "PC" represents parallel corpus based method, where the translation probability is given by the toolkit Giza++.

[3]http://www.shooter.cn. It is one of the most popular websites which provide subtitles with Chinese translations.

[4]http://ai.stanford.edu/ ehhuang/

[5]In this paper, we use dict.cn as the dictionary. It is one of the biggest online dictionary website in China.

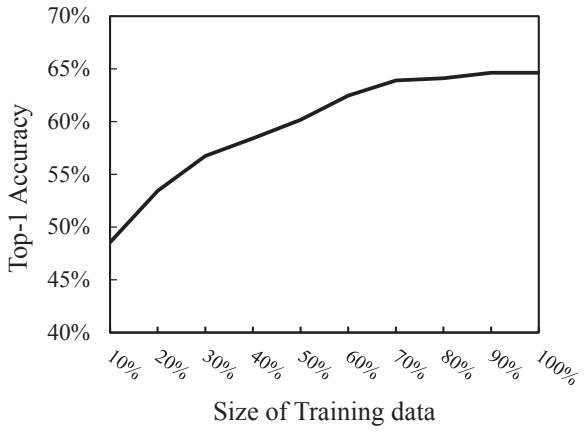

Figure 4: Sensitivity of translation probability obtained by Giza++ to number of training Data.

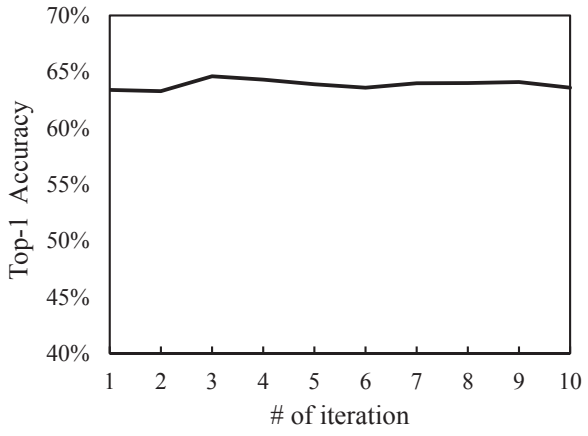

Figure 5: The impact of number of iterations of user aware neural network language model.

For language model, we compare the following methods:

- "LM" represents the traditional n-gram language model. In this work, we also use the toolkit Giza++ to train the model.
- "NLM" represents proposed by method proposed by Collobert and Weston [4].
- "NLM+U" represents the proposed user aware neural network language model.

Some existing works focused on constructing vocabularies for informal expressions can also be adapted for this task. In this work, we re-implemented works done by Han et al. [9] and Li and Yarowsky [16]. Since the string similarities between English tokens and their corresponding Chinese words are zero, the re-ranking part [9] is ignored. The window size is set to 3. Bi-gram is used to calculate the context similarity. For the method proposed by Li and Yarowsky [16], we used the data-driven hypothesis generation method and optimal weights in the log-liner model they used in their work. Vocabulary words generated based on parallel corpus are treated as the informal expressions.

The results of different word translation methods are shown in Table 2. From the results, we can observe that the proposed method, which incorporates noisy channel approach with neural network language model, achieve better performance than other methods. Comparing the results of dictionary based method with parallel corpus based method, we can see that the parallel corpus based method achieve better performance with all language models, although the two methods have the similar in-vocabulary words percentages. The main reason may be that movie and TV series have the similar language usages as social media. Comparing the results with different language models, we can observe that user aware neural network language model is better than neural network language model and n-gram language model. It demonstrates that the user historical information can benefit this task. Since the methods proposed by [16] and [9] mainly focus on frequent informal expressions, only English words with high frequency can be well processed by their work. However, these methods can process abbreviation, pronunciation, and some other kinds of English tokens.

For investigating the impact of training data used to estimate the word translation model, we showed the performance of PC+NLM+U with different number of training data for word translation model. Figure 4 shows the Top-1 accuracy with training data from 10% to 100% on the constructed parallel corpus. From the results, we can observe that the size of parallel corpus has a certain effect on the performance word translation model. We think that the increasing number of in-vocabulary words along with size of parallel corpus is one of main reasons.

To show the performance impact of the number of iterations of user aware neural network language model, we evaluate the accuracy of word translation method with different number of iterations. The results are shown in Fig. 5. From the figure, we can observe that it can achieve satisfactory accuracy with only one iteration. The top-1 accuracy achieved by user aware neural network language model with only one iteration is even better than the performance achieved by n-gram language models.

We also analyze the errors of the proposed method (PC-NLM+U) and find several types of words which cannot be well processed except the low frequency words. The first of these types are words which do not have Chinese translations in special context, for example, movie names, album names, etc. The second one are multiple words which should be translated as phrases. The third one are abbreviations which have multiple meanings. For example, "PE" can be used as the abbreviation of physical education, product engineering, private equity fund, and so on. We will leave these types of words for future work.

5.4 Word Categorization Results

Table 3 shows the word categorization results of different methods in five categories: person name, product name, organization name, informal expressions, and loanwords. We compare the following methods:

- "LP" represents the original label propagation without initialization method described in Eq.(12).
- "INIT" represents results of the method only based on Eq.(13).
- "INIT+LP/WS" denotes the proposed method with seed words.
- "INIT+LP/WOS" represents the proposed method without seed words.

Table 3: Word categorization results of different methods.

Methods	Person Name			Product Name			Org. Name			Slang			Loanwords			Acc.
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	
LP	37.5	23.8	29.1	**93.9**	18.4	30.8	**57.7**	17.4	26.8	19.2	31.3	23.8	8.8	**39.1**	14.4	22.5
INIT	23.1	**60.0**	**33.3**	36.2	27.6	31.3	28.1	52.3	36.6	27.9	45.8	34.6	42.9	13.0	20.0	32.3
INIT+LP/WOS	80.3	10.0	17.2	84.9	36.8	51.4	34.4	**89.5**	49.7	40.3	56.3	47.0	50.0	4.4	8.0	42.9
INIT+LP/WS	**90.9**	19.2	31.7	86.4	**58.6**	**69.8**	39.0	84.9	**53.4**	**48.7**	**75.0**	**59.0**	**57.1**	17.4	**26.7**	55.8

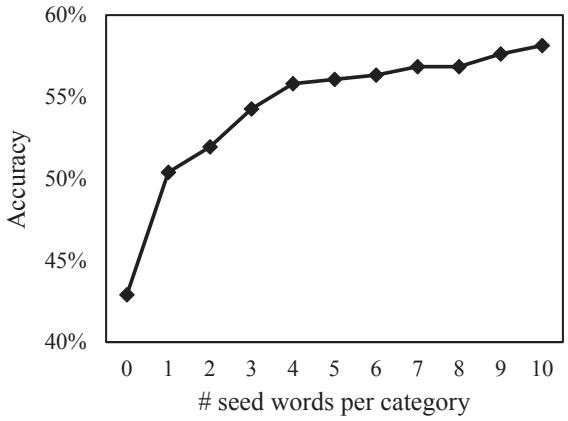

Figure 6: The impact of number of seed words per category.

For the methods INIT+LP/WS and LP, 4 seed words are used for each category. The total number of out-of-vocabulary words are 387, which takes about 32.3% percents among all English tokens.

From the results, we can observe that the propose method (INIT+LP/WS) achieves the best performances in four of the five categories. The accuracy of the proposed method is also significantly better than other methods. Comparing the results with and without the proposed initialization method (INIT+LP/WS v.s. LP), we can observe that initialization contributes a lot. The relative accuracy improvement is more than 148%. It demonstrates that the proposed method can achieve better performance with a small number of seed words.

From the Table 3, we also observe that all the methods cannot achieve satisfactory performances for the categories of person name and loanwords. We analyze the errors in these categories and find that lots of errors are caused by acronyms. Person names may be represented by acronyms. However, some of them are also used as the organization name. Since the number of webpages about a organization is usually larger than the number of webpages describing a person, most of them are classified into organization category. This is also one of the main reason of why all the methods achieve low precision in organization name category. These acronyms cannot be correctly classified without context information.

To show the performance impact of the number of seed words, we evaluate the accuracy of word categorization method with different number of seed words. The results are shown in Fig. 6. From the figure, we can observe that the accuracy is significantly improved with seed words comparing to the method without seed words. The accuracies are improving continuously along with the increasing the number of seed words. Although, when the number of seed words per category is more than 4 words, the performances increase slowly. We think that lack of context information is one of the main reasons.

Table 4: The perfomances of POS tagging of English tokens by different parsers with/without the proposed method. "WP" and "WoP" represent the accuracy with and without normalization method respectively.

Methods	WoP	WP
Stanford Parser 3.2	66.4%	84.7%
Berkeley Parser 1.7	67.9%	83.9%
FudanNLP 1.57	54.0%	79.6%

5.5 Applications

To show the effectiveness of the proposed method as the preprocessing step for other NLP tasks, we evaluate our method with three Chinese parsers. In this work, we only focus on the relative changes caused by English tokens. We randomly select 100 microblogs from the whole evaluation set. The in-vocabulary words are translated into Chinese. The words which cannot be translated are replaced by their category names.

Table 4 shows the accuracy of POS tagging of English tokens with and without the proposed normalization method. From the results, we can observe that all three parsers benefit a lot from the proposed normalization method in processing mixed texts. The relative improvements are significant. Since features extracted from words play important rules in existing methods, lack of word information may highly impact their performances. For the methods which label all English tokens with a same tag [24], the proposed method can bring much more benefit for them.

6. CONCLUSIONS

In this paper, we focus on the task of normalizing the Chinese-English mixed texts. We firstly analyzed the phenomena of mixed usage in Chinese. Then, we propose to use word translation and categorization to achieve the task. For word translation, we use noisy-channel approach with neural network language model to translate in-vocabulary words. A novel user aware neural network language model is introduced to capture the useful historical information of users. For categorizing words, a graph-based unsupervised method is proposed . We also introduce a novel initialization technique to improve the effectiveness. Experimental results show that the top-5 word translation accuracy of proposed method achieves 86.2%. For word categorization, the propose method also achieves significant improvement over the baseline methods. The relative improvement over the original label

propagation method is more than 148%. We also incorporate the proposed method as preprocessing steps for three different parsers. All of them benefit a lot from the proposed method.

7. ACKNOWLEDGEMENT

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (61003092, 61073069), National Major Science and Technology Special Project of China (2014ZX03006005), Shanghai Municipal Science and Technology Commission (No.12511504502), Key Projects in the National Science & Technology Pillar Program(2012BAH18B01) and "Chen Guang" project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation(11CG05).

8. REFERENCES

[1] A. Aw, M. Zhang, J. Xiao, and J. Su. A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40, Sydney, Australia, July 2006. Association for Computational Linguistics.

[2] R. Beaufort, S. Roekhaut, L.-A. Cougnon, and C. Fairon. A hybrid rule/model-based finite-state framework for normalizing sms messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 770–779, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[3] J.-S. Chang and W.-L. Teng. Mining atomic chinese abbreviations with a probabilistic single character recovery model. *Language Resources and Evaluation*, 40(3-4):367–374, 2006.

[4] R. Collobert and J. Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 160–167, New York, NY, USA, 2008. ACM.

[5] D. Das and S. Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[6] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics, 2013.

[7] D. Freitag. Machine learning for information extraction in informal domains. *Machine Learning*, 39(2-3):169–202, 2000.

[8] B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[9] B. Han, P. Cook, and T. Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 421–432, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[10] B. Han, P. Cook, and T. Baldwin. Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, 4(1):5:1–5:27, Feb. 2013.

[11] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[12] M. Johnson and A. E. Ural. Reranking the berkeley and brown parsers. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 665–668, Los Angeles, California, June 2010. Association for Computational Linguistics.

[13] C. Kobus, F. Yvon, and G. Damnati. Normalizing sms: are two metaphors better than one? In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 441–448, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[14] C. Li and Y. Liu. Improving text normalization using character-blocks based models and system combination. In *Proceedings of COLING 2012*, pages 1587–1602, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[15] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 339–346, New York, NY, USA, 2008. ACM.

[16] Z. Li and D. Yarowsky. Mining and modeling relations between formal and informal chinese phrases from web corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1031–1040, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[17] Z. Li and D. Yarowsky. Unsupervised translation induction for chinese abbreviations using monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 425–433, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[18] F. Liu, F. Weng, and X. Jiang. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 1035–1044, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[19] E. Minkov, R. C. Wang, and W. W. Cohen. Extracting personal names from email: applying named entity recognition to informal text. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 443–450, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[20] T. Mullen and R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[21] Z.-Y. Niu, D.-H. Ji, and C. L. Tan. Word sense disambiguation using label propagation based

semi-supervised learning. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 395–402, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[22] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, Mar. 2003.

[23] N. Okazaki, M. Ishizuka, and J. Tsujii. A discriminative approach to japanese abbreviation extraction. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 889–894, 2008.

[24] X. Qian, Q. Zhang, X. Huang, and L. Wu. 2d trie for fast parsing. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 904–912, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[25] X. Qiu, Q. Zhang, and X. Huang. Fudannlp: A toolkit for chinese natural language processing. In *Proceedings of ACL*, 2013.

[26] G. Richard. A global perspective on bilingualism and bilingual education. *Georgetown University Round Table on Languages and Linguistics 1999: Language in Our Time: Bilingual Education and Official English, Ebonics and Standard English, Immigration and the Unz Initiative Languages and Linguistics 1999*, page 332, 2001.

[27] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[28] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing with compositional vector grammars. In *Proceedings of ACL 2013*, June 2013.

[29] A. Tamura, T. Watanabe, and E. Sumita. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 24–36, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[30] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[31] P. D. Turney and M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *(No. ERB-1094, NRC #44929): National Research Council of Canada*, 2002.

[32] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 777–785, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[33] P. Wang and H. T. Ng. A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 471–481, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[34] L.-X. Xie, Y.-B. Zheng, Z.-Y. Liu, M.-S. Sun, and C.-H. Wang. Extracting chinese abbreviation-definition pairs from anchor texts. In *Machine Learning and Cybernetics (ICMLC)*, volume 4, pages 1485–1491, 2011.

[35] D. Yang, Y.-C. Pan, and S. Furui. Vocabulary expansion through automatic abbreviation generation for chinese voice search. *Computer Speech & Language*, 26(5):321 – 335, 2012.

[36] J. Zhao, X. Qiu, S. Zhang, F. Ji, and X. Huang. Part-of-speech tagging for chinese-english mixed texts with dynamic features. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1379–1388, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[37] X. Zhu and Z. Ghahramani. Learning from Labeled and Unlabeled Data with Label Propagation. In *Technical Report CMU-CALD-02-107*. Carnegie Mellon University, 2002.