# ADDRESSING INFORMALITY IN PROCESSING CHINESE MICROTEXT

**AOBO WANG**

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2014

# DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

AOBO WANG

31 December 2014

# ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to everyone who supported me throughout my doctoral candidature. First and most important of all, I want to thank my advisor A/P Min-Yen Kan for his supervision and help throughout this entire period of time. I am thankful for his aspiring guidance, invaluably constructive criticism and friendly advice during my doctoral candidature.

I express my warm thanks to members of my thesis committee, including Prof Chew-Lim Tan, Prof Khe-Chai SIM, and Prof Jong C. Park who have put in effort to review and assess this thesis.

I would also like to thank my wife, Jing Chen, and my family. Their patience and support is the key to accomplish my doctoral candidature.

I am sincerely grateful to many others who have shared their truthful

and illuminating views on my research topic. Here I would like to thank my friends and colleagues including Ziheng, Zhao Jin, Jesse, Jun Ping, Chen Tao, Xiangnan, Jovian, An Qi, Kai Ishikawa, Daniel Andrade and Takashi Onishi.

# CONTENTS

# ABSTRACT

In this thesis, I tackle the problem of processing Chinese microtext, with the goal of building the natural language processing (NLP) tools for the microtext domain. I discover that informal words and named entities that are formed in a free-style manner are key reasons why microtext is difficult to understand and process by conventional nature language processing tools. As such in this thesis, I study three key areas to address informality in processing Chinese microtext:

1. informal word recognition and word segmentation,
2. informal word normalization, and
3. named entity recognition.

The first area allows us to identify the unknown, informal words formed from ordinary Chinese characters, resulting in improved word segmentation. By leveraging my observation of the mutual dependence between informal word recognition and word segmentation, I formulate the problem as a two-layer sequential labeling problem for which a factorial conditional random field is used to perform both tasks jointly. This joint inference method significantly outperforms baseline systems that conduct the tasks individually or sequentially.

The second area links informal words to their formal counterparts, which can help both human and machine better understand these informal replacements. I formalize the task as a classification problem and propose rule-based and statistical features to model three plausible channels that explain the connection between formal/informal pairs. I evaluate my two-stage selection-classification model on a crowdsourced corpus, achieving a normalization precision of 89.5% across the different channels, significantly improving the state-of-the-art.

The third area targets the important class of words in common in microtext: named entities. I propose an effective method to obtain annotations for named entities automatically and employ the conditional random field to label named entities in microtext, using features derived from both labeled and unlabeled data. To further improve the performance gains derived from the automatic annotations, my method caters for the time-sensitivity nature of named entities, thus keeping the model up-to-date.

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1

# Introduction

User generated content (UGC) initially made its first appearance with the web log (blog). But over the past decade, it has since matured to allow participation by the masses with very little time commitment. Current trends point to short texts – including microblogs, comments, SMS, chat and instant messaging; collectively referred to as *microtext* by Gouws et al. (2011) or *network informal language* by Xia et al. (2005) – as the hallmark of the participatory Web. Unlike formal text (*i.e.*, newswire, textbook and scientific articles), the short, informal nature of microtext allows users to post frequently and quickly react to others posts, making social network platforms (*e.g.*, Twitter [1], Facebook [2] and Sina Weibo [3]) important forms of close-to-real-time communication. Research has also followed these shifting trends in content, recognizing UGC as a valuable source of data for downstream applications. Microtext has been widely studied as a source of data for information filtering (Diaz-Aviles et al., 2012; Sriram et al., 2010), sentiment analysis (Brody and Diakopoulos, 2011; Liu et al.,

[1] `https://twitter.com`

[2] http://www.facebook.com

[3] The most popular Chinese social media, akin to a hybrid of Twitter and Facebook.`http://open.weibo.com`

2012), classification (Nishida et al., 2012; Zubiaga et al., 2011), event discovery (Achananuparp et al., 2012; Song et al., 2012) and sharing (Fujiki et al., 2011; Wang et al., 2012a).

While a rich source that many applications are interested in mining for knowledge, microtext is difficult to process due to its informality. One key reason for this is the ubiquitous presence of informal words and free-styled named entities. Informal words refer to anomalous terms that are synonyms of their formal counterparts, but manifest as ad hoc abbreviations, neologisms, unconventional spellings and phonetic substitutions. Unlike their formal counterparts that are widely and easily understood, informal words are often not straightforward to interpret by the general public, yet frequently used among certain cliques in social media. Named entities in microtext further raise difficulties with their variety and free-style method for formation. These traits of microtext cause its natural language to evolve at an astonishing rate, far outstripping the pace of lexicographers updating dictionaries. Such informal characteristics frustrate Natural Language Processing (NLP) tools that have largely been trained and used in formal text (*i.e.*, newswire) domains. To address these problems, recent work has recognized this and started to address this performance gap Han and Baldwin (2011); Kobus et al. (2008); Xia and Wong (2006).

While the focus of microtext processing has been on English, it also is clear that these language trends are global, and have impacted microtext transcending language boundaries. It is important to develop non-English language tools as well as corpora to broaden analytics on microtext. Different from most previous studies (Beaufort et al., 2010; Kobus et al., 2008; Liu et al., 2011a) working on English, my work pays specific attention to address the challenges raised by Chinese microtext. In the remaining part of this chapter, I will motivate my work and summarize the key contributions I have achieved.

## 1.1 Informal Words in Chinese Microtext

Informal words, referred to as network informal language (NIL) expressions by Xia et al. (2005), initially drew researchers' attentions with their first public appearance in bulletin board system (BBS) chats. As microtext is often timely, its production follows social trends and news events in social media, and thus its informal words and their usage evolves rapidly, causing NLP tools to fail. For example, considering the microtext listed in Figure 1.1: "开发区木有出租车", a machine translation system may mistranslate it literally as "There are taxis in the development zone", by ignoring the "weird" word "木"("wood" is its literal meaning.) This occurs as the translation system may not know the informal word "木有" ("没有" ; "no"). It is thus desirable to pay special attention on informal words before proceeding with typical text processing workflows.

To bootstrap my study, I utilize the Chinese social media archive, PrEV (Cui et al., 2012), to obtain Chinese microblog posts from the public timeline of Sina Weibo[4]. This *Dataset A* has a total of 6,678,021 messages, covering two months from June to July of 2011. To annotate the corpus, I employ *Zhubajie*[5], one of China mainland's largest crowdsourcing (Wang et al., 2010) platforms to obtain necessary annotations, as detailed in later chapters.

[4] http://open.weibo.com

[5] http://www.zhubajie.com

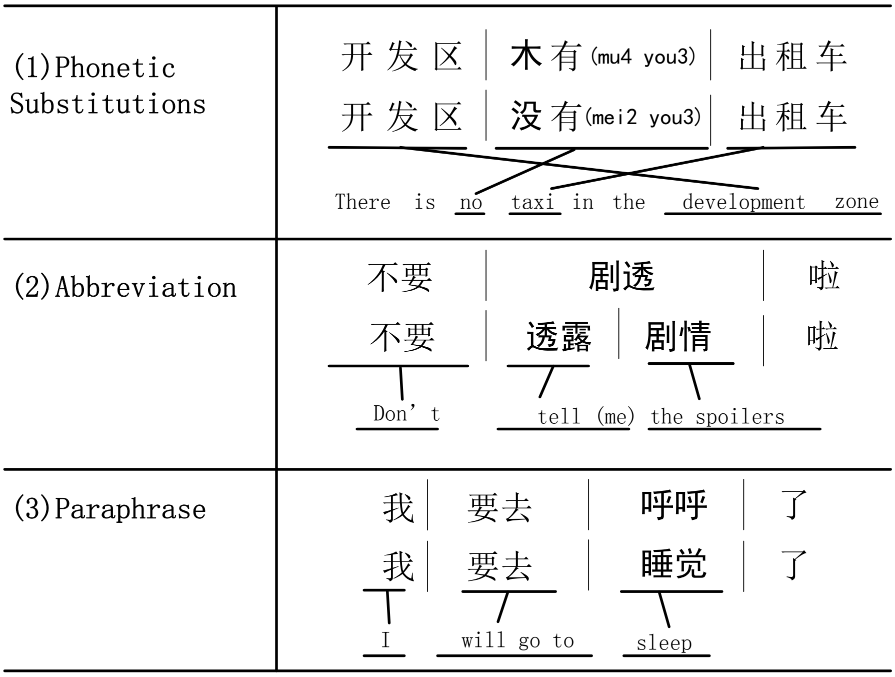


Figure 1.1: The classification of informal words. The three layers give informal words (bolded) with example sentences, the normalized Chinese form and aligned English translation.

### 1.1.1 Informal Word Recognition and Word Segmentation

In particular, the recognition of informal words is an important pre-processing step for NLP tasks that rely on keyword matching or word frequency statistics. Given this example tweet: "Toooo tired for life", the informal word "Toooo" can be easily recognized by a dictionary-based method. But unlike such noisy words in English, Chinese informal words are more difficult to mechanically recognize due to two critical reasons: first, Chinese does not employ word delimiters; second, Chinese informal words combine numbers, alphabetic letters, Chinese characters and even punctuation. Techniques for English informal word detection that rely on word boundaries and informal word orthography do not work in Chinese, and thus the problem needs to be approached quite differently for Chinese.

Consider the microtext “不要剧透了” (meaning “Don’t tell me the spoilers (to a movie or joke)”) in Figure 1.1. If “不要” (“don’t”) and “ 了” (past tense marker) are correctly recognized as two words, we may predict the previously unseen characters “剧透” (“tell spoilers”) as an informal word, based on the learned Chinese language patterns. However, state-of-the-art Chinese segmenters[6] incorrectly yield “不要␣剧␣透了”, preferring to chunk “透了” (“thoroughly”) as one word, since they do not consider the possibility that “剧透” (“spoiler”) could be an informal word.

The example highlights the strong dependency between adjacent words, where the ignorance of informal words lead to incorrect segmentation. Moreover, we also know that correct word segmentation performance can ease informal word recognition. Which problem should be tackled first? Might there be a method that can capitalize on the intertwined nature of these two problems?

The answer is “yes”. Chapter 3 shows that such synergy between Chinese word segmentation (CWS) and informal word recognition (IWR) can be exploited through joint inference. Rather than pipeline the two processes serially, I recognize that the two problems have a strong mutual dependency that should be solved jointly. I formulate the problem as a two-layer sequential labeling problem for which a factorial conditional random field (FCRF) is used to perform both CWS and IWR.

Joint CWS and IWR can help downstream Chinese microtext natural language processing, however, it also intrinsically has value. From a lexicographical standpoint, IWR can help build tools to recognize, capture and codify informal words, akin to sites like Wordnik [7] for English. I build a Chinese informal word lexicon by automatically analyzing over 6 million

[6] http://www.ictclas.org/index.html

[7] www.wordnik.com

Weibo microblog posts, leveraging our dynamic CRF approach. The Web-based lexicon shows a concordance view of example microblog posts where the potential informal word occurs. It shows that over 50% of the identified words in our ground-truth portion were not recorded in the well-known collaborative Chinese encyclopedia Baidu Baike [8], as of the time my research was being conducted.

### 1.1.2 Informal Word Normalization

Given the close connection between an informal word and its formal equivalent, the restoration (normalization) of an informal word to its formal counterpart is the next important process that typically follows IWR. Taking the same tweet " Toooo tired for life" as an example, following recognition, the next step is to restore "Toooo" as "too" so that it can be properly recognized as an intensifier for downstream analysis (*e.g.*, sentiment analysis). A series of previous work (Beaufort et al., 2010; Kobus et al., 2008; Liu et al., 2011a) have addressed this normalization issue for English microtext. These methods are promising in mining English formal counterparts from a dictionary based on the morphologic similarity and edit distance between formal/informal pairs, however they can not be directly adopted to fully address the informal word normalization (IWN) problem on Chinese microtext.

First of all, there is no trivial way to generate formal candidate word list. Moreover, the connections between informal/formal pairs are not limited to phonetic or morphologic similarity. As shown in Figure 1.1, the pairs in **Paraphrase** channel have similarity in semantic meaning and character usage, but neither in pronunciation nor spelling. Kobus et al. (2008) and

[8] `www.baike.baidu.com`

Aw et al. (2006) treated the informal and formal text as two different languages, and adopted phrase-based machine translation (MT) model to perform microtext normalization. The MT-like method requires a relatively large number of manually normalized informal/formal sentences pairs as training data, which can be expensive to obtain.

In Chapter 4, I present a novel method for normalizing informal word to their formal equivalents. Specifically, given an informal word with its context as input, I generate hypotheses for its formal equivalents by searching the Google Web 1T corpus[9] (Brants and Franz, 2006). Prospective informal/formal pairs are further classified by a supervised binary classifier to identify correct pairs. In the classification model, I incorporate both rule-based and statistical feature functions that are learned from both gold-standard annotation as well as formal domain synonym dictionaries. Also importantly, this method does not directly use plain words or lexica as features, keeping the learned model small yet robust to inevitable vocabulary change.

## 1.2 Named Entity Recognition on Microtext

Named entity recognition (NER) in formal texts has been studied through several distinct communities through shared tasks, (such as MUC [10], CONLL [11], ACE [12] and SIGHAN [13]), in which *Person*, *Organization* and *Location* are studied as the three typical types of named entities (NEs). Are these the same named entities that occur in microtext? To further characterize the named entities in microtext, Three annotators manually labeled 1000 mi-

[9] `https://catalog.ldc.upenn.edu/LDC2006T13`

[10] `http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html`

[11] `http://www.clips.ua.ac.be/conll2002/`

[12] `http://www.itl.nist.gov/iad/mig/tests/ace/`

[13] `http://www.sighan.org/bakeoff2006/`

crotext posts (from the portion gathered during July 2011), gathering a set of 883 unique named entities (hereafter, *Dataset B*), and compared it with another NE list from the MSRA corpus (in the formal domain) released in SIGHAN shared task. Table 1.1 shows the major types of named entities with their distributions, as well as the average count of words per named entity in both domains.

I make two observations from this.

Firstly, it suggests that in addition to the conventional types of NEs, *Nickname*, *Title* and *Product* are frequently mentioned in microtext hence deserving more research attention. I note that although *Nickname* is a frequent NE type in microtext, the functional tag "@" that prefixes nickname use eases its recognition significantly. I thus focus my attention on *Title* (*i.e.*, titles of the books, movies, TV shows, games, etc.) and *Product* (*i.e.*, general objects offered to a market), which are hot topics in microtext as sharing and commenting on these topics are major social media activities.

Table 1.1: The distribution of different types of NEs. **AvgLenIF** (**AvgLenF**) refers to the average count of words per NE in informal (formal) domain.

| **NEs Type** | **Percentage** (%) | **AvgLenIF** | **AvgLenF** |
|---|---|---|---|
| **Person** | 10.8 | 1.10 | 1.04 |
| **Location** | 10.7 | 1.31 | 1.08 |
| **Organization** | 14.1 | 1.49 | 1.56 |
| **Nickname** | 27.0 | 2.69 | — |
| **Title** | 17.9 | 2.67 | — |
| **Product** | 8.0 | 1.91 | — |

Secondly, *Title* and *Product* are largely made by free-styled combination of characters, words and alphabetic letters, different from typical single-word NEs common in formal text. For example, in Figure 1.2, unlike the conventional types of NEs that contain high indicator words or characters (*e.g.*, standard Chinese surnames like “张”, “王”,“李”; regular location names like “中”[“China”], “美”[“the US”]; and common suffixes such as “公司”[“company”], “局”[“office”], “部”[“department”]), informal microtext NEs exhibit diverse vocabulary and few indicators, increasing the sparsity of lexical features. Confounding this, most component words of informal NEs (*e.g.*“宝石” [“jewelry”], “看” [“watch”],“手机” [“cellphone”]) are also frequently used as non-NE words in text, further raising ambiguity.

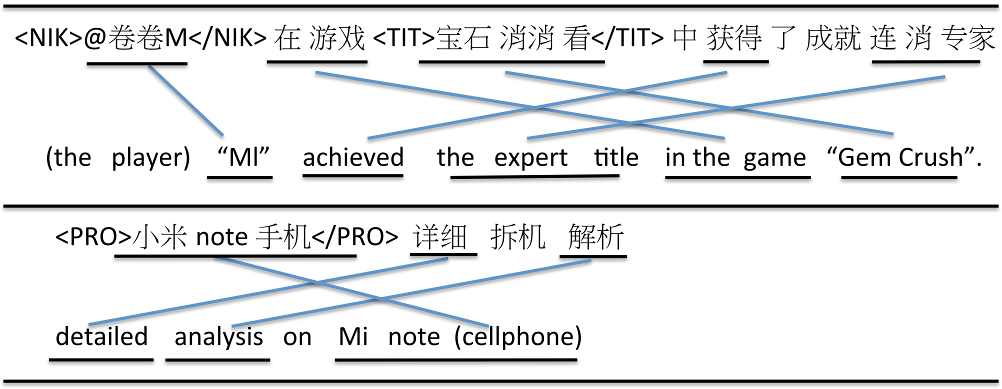

<NIK>@卷卷M</NIK> 在 游戏 <TIT>宝石 消消 看</TIT> 中 获得 了 成就 连 消 专家
(the player) “MI” achieved the expert title in the game “Gem Crush”.
<PRO>小米 note 手机</PRO> 详细 拆机 解析
detailed analysis on Mi note (cellphone)

Figure 1.2: Example Chinese microtext with NE annotation and aligned translation. “NIK”,“TIT” and “PRO” refer to the label for *Nickname Title* and *Product*

Furthermore, the lack of training corpus for the microtext domain is also a bottleneck for supervised strategies. I have already illustrated the great domain mismatch between formal and informal text – this can easily frustrate the conventional recognizers trained on formal text domain. However, due to the fairweather nature of trending topics on social media, even an annotated corpus may quickly become obsolescent. It is essential to tune NER in microtext to account for its time-sensitive nature.

To fill these gaps, I propose an effective method to obtain annotations automatically and employ the conditional random field model to label the *Title* and *Product* in microtext, using features derived from both labeled and unlabeled data. Time-sensitivity is also incorporated to keep the model up-to-date and within a reasonable size limit.

## 1.3 Key Contributions

In brief, my thesis studies the processing of Chinese microtext via three key areas: 1) informal word recognition and word segmentation, 2) informal word normalization, and 3) named entity recognition

The key contributions of this thesis include:

1. Leveraging the dependency between CWS and IWR, I propose the usage of factorial conditional random field model to jointly tackle the two problems, achieving significant improvement in performance over the state-of-the-art on both tasks (Wang and Kan, 2013),

2. To operationalize informal word normalization, I suggest a novel two-stage candidate generation-classification method, bettering the current state of the art with respect to both $F_1$ and loss rate (Wang et al., 2013),

3. Having built the Chinese microtext corpus [14] with crowdsourced annotations, I perform a systematic analysis on the informal words origin, sentiment and usage (Wang and Kan, 2013; Wang et al., 2013), and

4. I present an effective method to recognize the time-sensitive named entities in microtext with the corpus crawled and annotated automatically.

[14] The dataset is publicly available at `http://wing.comp.nus.edu.sg/downloads/weiboCWSIWRData/`

## 1.4 Organization

In the next chapter, I will provide a detailed literature review on related research areas. Then in Chapter 3, I will explain the work that I have done for Chinese word segmentation and formal word recognition. Chapter 4 presents my work for informal word normalization. Chapter 5 presents my method for recognizing named entities from raw microtext. The last chapter concludes this thesis, highlighting opportunities for further research.

# Chapter 2

# Related Work

Having motivated the thesis in the previous chapter, I review the relevant related work in the areas of 1) Chinese word segmentation, 2) Informal Chinese words recognition and normalization, and 3) Named entity recognition. Through this, I detail the developments of these respective fields with a focus on the domain of microtext in social media.

In the first section, I start with the Chinese word segmentation over two milestones: 1) related works on Chinese word segmentation, and 2) the typical enhancement towards new word recognition. I continue the review with a new section discussing the related work on informal word recognition and normalization, which is one of the critical research problems studied by this thesis. Moving on to named entity recognition, I present a brief summary on the technologies designed for conventional types of NEs and highlight the key achievements in shared tasks on Chinese NER. Finally, to conclude my review of related work, I discuss the the research effort on tackling the domain mismatch problem, which is to recognize named entities from microtext.

## 2.1 Chinese Word Segmentation

CWS has been widely recognized as a preliminary and crucial pre-process for Chinese language processing, as Chinese lacks word delimiters unlike most Western languages. During the last decade, from dictionary-based systems, rule-based systems to supervised machine-learning based systems, segmentation performance has been improved significantly, thanks to the considerable effort committed by the NLP community. After taking over the baton from support vector machine (*e.g.* Asahara et al. (2003); Li et al. (2005)) and maximum entropy (*e.g.* Low et al. (2005); Peng and Schuurmans (2001)) approaches, linear statistical models designed for sequential labeling problem (*e.g.*, Hidden Markov Model (HMM), Conditional Random Field (CRF) and their variants) have become dominant in the word segmentation market, with support from carefully-designed features (*e.g.*, bigram features, punctuation information (Li and Sun, 2009) and statistical information (Sun and Xu, 2011)). Specifically, during 2003, 2005 and 2010, word segmentation shared tasks in the SIGHAN workshop guided and shaped much of the research work, particularly by providing high quality annotated corpus and standardizing the evaluation approach. According to Huang and Zhao (2007), the evaluation in terms of such bakeoff data shows that the accuracy drop caused by out-of-vocabulary (OOV) words is at least five times greater than that of segmentation ambiguities. Better performance of OOV recognition offers the higher segmentation accuracy on the whole. Another key conclusion from the shared tasks is that the accuracy of statistical character-based tagging approaches dominate performance of any other word-based system.

Accompanying the development of word segmentation and especially the key finding of the importance of OOV, the community has put forth additional effort on Chinese new word detection. The task of Chinese new word detection is normally treated as a separate process from word segmentation in most previous works (Chen and Bai, 1998; Chen and Ma, 2002; Gao et al., 2005; Wu and Jiang, 2000). It is not surprising that researchers further propose to integrate the two tasks to benefit both. Work by Peng et al. (2004) and Sun et al. (2012) conducted segmentation and detection sequentially, but in an iterative manner rather than joint. This is a weakness as their linear CRF model requires re-training. Their method also requires thresholds to be set through heuristic tuning, as to whether the segmented words are indeed new words. I also highlight that the task of new word detection refers to OOV detection, and is distinctly different from informal word recognition (new words could be both formal or informal words), although these two are close.

## 2.2 Informal Word Recognition and Normalization

With English tweets, IWR has typically been investigated alongside normalization. Several recent works (Gouws et al., 2011; Han and Baldwin, 2011; Han et al., 2012) have aimed to produce informal/formal word lexica and mappings. These works are based on assumptions of distributional and string similarity that address concerns of lexical variation and spelling. These methods generally proposed a two-step unsupervised approaches to first detect and then normalize detected informal words using dictionaries.

### 2.2.1 Chinese Informal Word Recognition

In processing Chinese informal language, work conducted by Xia et al. (2005) addressed the problem of in bulletin board system (BBS) chats. They employed pattern matching and SVM-based classification to recognize Chinese informal sentences (not individual words) from chats. Both methods have their advantages: the learning-based method did better on recall, while the pattern matching performed better on precision. To obtain consistent performance on new unseen data, they further employed an error-driven method which performed more consistently over time-varying data (Xia and Wong, 2006). In contrast, my work identifies individual informal words, a finer-grained (and more difficult) task.

While seminal, I feel that the difference in scope (informal sentence detection rather than word detection) shows the limitation of their work for microblog IWR. Their chats cover only 651 unique informal words, as opposed to my study covering almost triple the word types ($1,658$). In my study of microtext, *Dataset A* demonstrates a higher ratio of informal word use (a new informal word appears in $\frac{1{,}658}{12{,}446} = 13\%$ of sentences, as opposed to $\frac{651}{22{,}400} = 2\%$ in their BBS corpus). Further analysis of their corpus reveals that phonetic substitution is the primary origin of informal words in their corpus – 99.2% as reported in Wong and Xia (2008). In contrast, the origin for informal words in microblogs is more varied, where phonetic substitutions abbreviations and neologisms, account for 53.1%, 21.4% and 18.7% of the informal word types, respectively. Their method is best suited for phonetic substitution, thus performing well on their corpus but poorly on the *Dataset A*.

Inspired by Wang et al. (2012c), who modeled the dependency between sentence boundary and punctuation jointly, in Chapter 3, I propose the usage of factorial conditional random field model to jointly tackle the CWS and IWR problems.

### 2.2.2 Chinese Informal Word Normalization

Normalization is a task that has much related work. Of particular relevance are the tasks of abbreviation normalization (*e.g.*, Chang and Teng (2006); Li and Yarowsky (2008b); Park and Byrd (2001); Zhang et al. (2011)), abbreviations and acronyms in medical domain Pakhomov (2002), and transliteration (*e.g.*, Bhargava and Kondrak (2011); Wu and Chang (2007); Zhang et al. (2010)). The basic idea behind these works is leveraging the edit distance to measure the similarity between acronyms and expansions. It is worth to note that simply re-training models trained on formal text or annotated microtext is insufficient: user-generated microtexts exhibit markedly different orthographic and syntactic constraints compared to their formal equivalents. For example, consider the microtext "河蟹社会" (formally, "和谐社会";"harmonious society"). A machine translation system may mistranslate it literally as "crab community" based on the meaning of its component words, if it lacks knowledge of the informal word "河蟹" ("和谐" ; "harmonious").

More closely related, Li and Yarowsky (2008a) tackled Chinese IWR and IWN in the Web domain. They bootstrapped 500 informal/formal word pairs by using manually-tuned queries to search for definition sentences – sentences that define or explain Chinese informal words with formal ones – through the use of a search engine[1]. The resulting noisy list was further reranked using a conditional log-linear model based on n-gram co-occurrence. However, their method makes a basic assumption that informal/formal

[1] www.baidu.com

word pairs co-occur within a definition sentence (i.e., "<*informal word*> means <*formal word*>"). This observation largely does not hold true in microblog data, as microbloggers largely do not define or explain the words they use. In addition, the features they proposed are limited to rule-based features and $n$-gram frequency, which does not permit their system to explain how the informal/formal word pair is related (*i.e.*, derived by which channel).

Wang and Ng (2013) analyzed 200 Chinese messages from Weibo and 200 English SMS messages from the NUS SMS corpus (Chen and Kan, 2011). Their analysis revealed that most incorrectly-spelled words were derived from correctly-spelled equivalents based on pronunciation similarity. In work on Chinese, this is often done by measuring the *pinyin* similarity between an informal/formal pair. (Li and Yarowsky, 2008a) computed the Levenshtein distance ($LD$) on the *pinyin* of the two words in the pair to reflect the phonetic similarity. However, as a general string metric, $LD$ does not capture the (dis-)similarity between two *pinyin* pronunciations well as it is too coarse-grained. To overcome this shortcoming, Wong and Xia (2008) proposed a source channel model that was extended with phonetic mapping rules and weights. They evaluated the model on manually-annotated phonetically similar informal/formal pairs. The disadvantage is that these rules and weights need to be manually created and tuned. In my work, the labor of manually creating rules and tuning weights is avoid, given annotated informal/formal pairs.

Furthermore, I make the key observation that the similarity of initial and final pairs are not independent, but may vary contextually. I will leverage this observation to make better use of the available annotations to incorporate both rule-based and statistical feature functions, in my proposed method describe in Chapter 4.

## 2.3 Chinese Named Entity Recognition

Conventional NER systems can be roughly divided into rule-based systems (Chiticariu et al., 2010; Sekine, 2004) and statistical machine learning systems. SVM (Kazama et al., 2002; Mayfield et al., 2003), HMM (Klein et al., 2003; Zhou and Su, 2002) and CRFs with its variants (Benajiba et al., 2010; Dinarelli and Rosset, 2011; Finkel and Manning, 2009, 2010; Seker and Eryigit, 2012). are the most popular statistical models for NER. Aside from the standard lexical and syntax features derived from external sources (*e.g.*, FreeBase[2]), parallel data (Benajiba et al., 2010; Munro and Manning, 2012), bilingual data (Che et al., 2013; Kim et al., 2012), as well as annotated parse trees (Finkel and Manning, 2009) have been proved to be efficient to tackle the NER tasks on the corpus with rich linguistic annotations. Current NER works on Chinese language (*e.g.*, Cheng et al. (2012); Du et al. (2010); Duan and Zheng (2011); Wang et al. (2012b)) mainly focus on recognizing single-word NEs from formal text such as news articles. The solid performance (Table 2.1) reported from SIGHAN[3] shared tasks suggests the promising ability of statistical models with abundant training data.

However, the focus of the above studies is limited to the recognition of the three major types of NEs, namely *Person*, *Organization* and *Location*. Named entities prevalent in the microtext domain have not been surveyed and are not addressed by current published work. To adapt such approaches for the microtext domain is difficult, as the critical bottlenecks of the lack of

[2] http://www.freebase.com/

[3] http://www.sighan.org/bakeoff2006/

Table 2.1: The best performance of NER systems in SIGHAN shared task. CityU and MSRA refer to the two datasets. "o" and "c" refer to the open and closed tracks, respectively.

| **Dataset** | **Pre**(%) | **Rec**(%) | $\mathbf{F}_1$(%) |
|---|---|---|---|
| CityU$_o$ | 93.42 | 87.43 | 90.33 |
| CityU$_c$ | 87.68 | 82.47 | 84.99 |
| MSRA$_o$ | 99.82 | 99.95 | 99.88 |
| MSRA$_c$ | 93.77 | 91.86 | 92.81 |

available large-scale training data and noise from automatically extracted linguistic features provided by other NLP tools (*e.g.*, POS tagger and syntax parser) cause significant performance degradation. Given the domain mismatch, current systems trained on formal text domain perform poorly on tweets. According to Liu et al. (2011b), the average $F_1$ of the Stanford NER (Finkel et al., 2005), which is trained on the CoNLL03[4] shared task data set, drops from 90.8% (Ratinov and Roth, 2009) to 45.8% on tweets.

I now review domain adaptation, as a pertinent method to address informal text is to consider it as adapting tools (here, specifically, NER) designed for formal texts for the informal domain. To tackle the topic adaptation problem formal text domain, bootstrapping via semi-supervised learning has been employed in several studies (*e.g.*, Jiang and Zhai (2007); Wu et al. (2009)). However, given the domain mismatch, the dataset from informal text is difficult to be labeled confidently and accurately, thus not reliable to be selected as bootstrapped instances. Recently Liu et al. (2011b) tackled named entity recognition on English tweets (i.e., microtext) using a semi-supervised approach by repeatedly retraining their K-Nearest Neighbors (KNN) and CRF model with an incrementally augmented training set. They sidestepped direct domain adaptation by using the annotated tweets, but the iterative re-training was time consuming.

[4] `http://www.cnts.ua.ac.be/conll2003/ner/`

In addition to bootstrapping, Ritter et al. (2011) demonstrated the utility of linguistic features provided by pre-trained Twitter specific POS taggers and shallow parsers. Although the linguistic features is proved to be valuable, the annotation for POS tagging and shallow parsers is relatively expensive and in their work, limited to 800 tweets. Hence it is desirable to utilize strategies that eschew expensive annotations. One possible solution proposed by Li et al. (2012) is a two-step unsupervised NER system specially designed for "targeted" Twitter stream. In the first step, they leveraged on Wikipedia and Microsoft Web N-Gram corpus to partition tweets into segments (candidate NEs), using a dynamic programming algorithm. In the second step, a random walk model is constructed to rank the segments by the probability of being true named entities. The final performance largely depends on multiple untuned parameters in the ranking model, which could be further improved. In a parallel development, Wan et al. (2011) conducted the study on recognizing NEs in Chinese news comments, also a form user-generated microtext. They leveraged the entity information in the original news article to revise the outputs of a linear CRF based NER that performed sequence labeling on the corresponding comments. This idea is reasonable as the news and comments are naturally related, but not a pertinent strategy for general microtext, since linking microblog posts to relative news articles is still challenging. In other related work on product reviews crawled from Amazon.com, Min and Park (2012) paid attention to classify annotated product names into temporal classes. They also conducted an experiment on product name extraction based on a parser with a regular grammar (Bird, 2006), together with predefined product name patterns, which are limited to English reviews.

In my work, I have proposed an effective method to obtain annotations automatically and employed CRF model to label the*Title* and *Product* in microtext. I am also able to keep the model up-to-date by considering the time-sensitivity of NEs.

## 2.4 Summary

This chapter reviews the related work of the thesis including word segmentation, informal word recognition and normalization, as well as named entity recognition on microtext. In the following chapters, I will detail the methodologies proposed to each of these problems motivated and the evaluation through experiments.

# Chapter 3

# Joint Word Recognition and Segmentation on Chinese Microtext

For the first key area, I focus on the informal word recognition problem. I make the key observation that Chinese word segmentation and informal word recognition are intricately related. Up until now, previous work has taken microtext processing to be a post-processing step after word recognition. However for Chinese, word segmentation is itself a non-trivial processing task, and is particularly problematic for informal microtext. It makes sense to consider these two problems together. To address the problem, my method simultaneously segments the input Chinese microtext into words and marks the component words as informal or formal ones.

## 3.1 Joint Inference

### 3.1.1 Problem Formalization

The two tasks are simple to formalize. The IWR task labels each Chinese character with either an **F** (part of a formal word) or **IF** (informal word). For the CWS task, I follow the widely-used **BIES** coding scheme (Hai et al., 2006; Low et al., 2005), where **B**, **I**, **E** and **S** stand for *beginning of a word*, *inside a word*, *end of a word* and *single-character word*, respectively. As a result, there are two (hidden) labels to associate with each (observable) character. Figure 3.1 illustrates an example microblog post graphically, where the labels are in circles and the observations are in squares. The two informal words in the example post are "木有" (normalized form: "没有"; English gloss: "no") and "rp" ("人品值"; "luck").

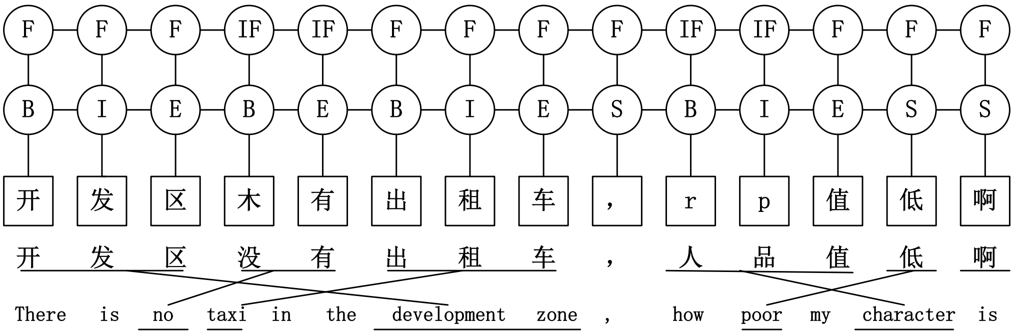

F F F IF IF F F F F IF IF F F F
B I E B E B I E S B I E S S
开 发 区 木 有 出 租 车 ， r p 值 低 啊
开 发 区 没 有 出 租 车 ， 人 品 值 低 啊
There is no taxi in the development zone , how poor my character is

Figure 3.1: A Chinese microtext (bottom layer) with annotations for IWR (top layer) and CWS (middle layer). The bottom three lines give the normalized Chinese form, its pronunciation in *pinyin* and aligned English translation.

### 3.1.2 Conditional Random Field Models

Given its general performance and discriminative framework, Conditional Random Fields (CRFs) (Lafferty et al., 2001) is a suitable framework for tackling sequence labeling problems. Other alternative frameworks such as

Markov Logic Networks (MLNs) and Integer Linear Programming (ILP) could also be considered. However, for this task formulating efficient global formulas (constraints) for MLN (ILP) is less straightforward than in other tasks (*e.g.*, compared to Semantic Role Labeling, where the rules may come directly from grammatical constraints). CRFs represent a basic, simple and well-understood framework for sequence labeling, making it a suitable framework for adapting to perform joint inference.

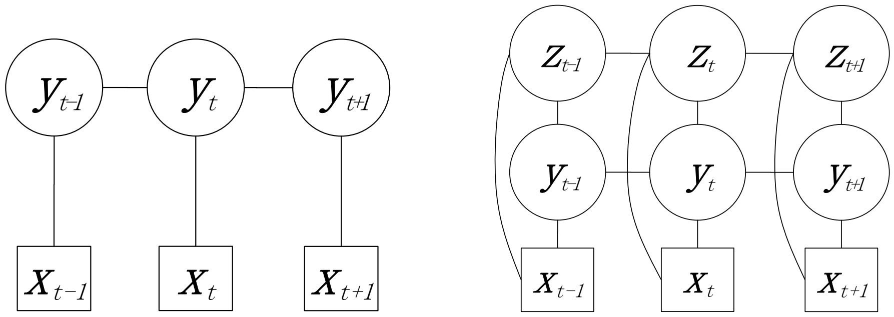
### 3.1.3 Linear-Chain CRF

A linear-chain CRF (LCRF; Figure 3.2a) predicts the output label based on feature functions provided by the scientist on the input. In fact, the LCRF has been used for the exact problem of CWS (Sun and Xu, 2011), garnering state-of-the-art performance, and as such, validate it as a strong baseline for comparison.

### 3.1.4 Factorial CRF

To properly model the interplay between the two sub-problems, I employ the factorial CRF (FCRF) model, which is based on the Dynamic CRF (DCRF) (Sutton et al., 2007). By introducing a pairwise factor between different variables at each position, the FCRF model results as a special case of the DCRF. A FCRF model captures the joint distribution among various layers and jointly predicts across layers. Figure 3.2 illustrates both the LCRF and FCRF models, where cliques include within-chain edges (e.g., $y_t$, $y_{t+1}$) in both LCRF and FCRF models, and the between-chain edges (e.g., $y_t$, $z_t$) only in the FCRF.

Although the FCRF can be collapsed into a LCRF whose state space is the cross-product of the outcomes of the state variables (i.e., 8 labels in



(a) Linear-chain CRF (b) Two-layer Factorial CRF

Figure 3.2: Graphical representations of the two types of CRFs used in this work. $y_t$ denotes the $1^{st}$ layer label, $z_t$ denotes the $2^{nd}$ layer label, and $x_t$ denotes the observation sequence.

this case), Sutton et al. (2007) noted that such a LCRF requires not only more parameters in the number of variables, but also more training data to achieve equivalent performance with an FCRF. Given the limited scale of the state space and training data, I follow the FCRF model, using exact Junction Tree (Jensen, 1996) inference and decoding algorithm to perform prediction.

### 3.1.5 CRF Features

I use three broad feature classes – lexical, dictionary-based and statistical features – aiming to distinguish the output classes for the CWS and IRW problems. Character-based sequence labeling is employed for word segmentation due to its simplicity and robustness to the unknown word problem (Xue, 2003).

A key contribution of this work is also to propose novel features for joint inference. I propose new features for the dictionary-based and statistical feature classes, which I have marked in the discussion below with "(*)". I later examine their efficacy in Section 3.2.4.

**Lexical Features**. As a foundation, I employ lexical (n-gram) features informed by the previous state-of-the-art for CWS (Low et al., 2005; Sun and Xu, 2011). These features are listed below[1]:

- Character 1-gram: $C_k(i-4<k<i+4)$
- Character 2-gram: $C_kC_{k+1}(i-4<k<i+3)$
- Character 3-gram: $C_kC_{k+1}C_{k+2}(i-3<k<i+2)$
- Character lexicon: $C_{-1}C_1$

  This feature is used to capture the common indicators in Chinese interrogative sentences. (e.g., "是不是" ("whether or not"), "好不好" ("OK or not"))
- Whether $C_k$ and $C_{k+1}$ are identical, for $i-4<k<i+3$.

  This feature is used to capture the words of employing character doubling in Chinese. (e.g., "拜拜" ("see you"), "天天" ("every day"))

**Dictionary-based Features**. I use features that indicate whether the input character sequence matches entries in certain lexica. I use an online dictionary from Peking University as the formal lexicon and the compiled informal word list from the training instances as the informal lexicon. In addition, I employ additional online word lists[2] to distinguish named entities and function words from potential informal words.

Alphabetic sequences in microblogs may also refer to Chinese *pinyin* or *pinyin* abbreviations, rather than English (e.g., "*bs*" for "鄙视 *bi shi*"; "to despise"). Hence, I added dictionary-based features to indicate the presence of *pinyin* initials, finals and standard *pinyin* expansions, using a

[1] For notational convenience, I denote a candidate character token $C_i$ as having a context $...C_{i-1}C_iC_{i+1}...$. I use $C_{m:n}$ to express a subsequence starting at the position $m$ and ending at $n$. $len$ stands for the length of the subsequence, and $offset$ denotes the position offset of $C_{m:n}$ from the current character $C_i$. I use $b$ (*beginning*), $m$ (*middle*) and $e$ (*ending*) to indicate the position of $C_k$ ($m \leq k \leq n$) within the string $C_{m:n}$.

[2] Resources are available at `http://www.sogou.com/labs/resources.html`, accessible as of 18 December 2014

UK English word list[3]. The final list of dictionary-based features employed are:

- If $C_k\,(i-4<k<i+4)$ is a surname: *Surname@k*
- (*) If $C_k\,(i-4<k<i+4)$ is a stop word: *StopW@k*
- (*) If $C_k\,(i-4<k<i+4)$ is a noun-suffix: *NSuffix@k*
- (*) If $C_k\,(i-4<k<i+4)$ is a *pinyin* Initial: *Initial@k*
- (*) If $C_k\,(i-4<k<i+4)$ is a *pinyin* Final: *Final@k*
- If $C_k\,(i-4<k<i+4)$ is a English letter: *En@k*
- If $C_{m:n}\,(i-4<m<n<i+4, 0<n-m<5)$ matches one entry in the Peking University dictionary:

  *FW@m:n*; *len@offset*; *FW-*$C_k$*@b-offset*, *FW-*$C_k$*@n-offset* or *FW-*$C_k$*@e-offset*
- (*) If $C_{m:n}\,(i-4<m<n<i+4, 0<n-m<5)$ matches one entry in the informal word list:

  *IFW@m:n*; *len@offset*; *IFW-*$C_k$*@b-offset*, *IFW-*$C_k$*@n-offset* or *IFW-*$C_k$*@e-offset*
- (*) If $C_{m:n}\,(i-4<m<n<i+4, 0<n-m<5)$ matches one entry in the valid *pinyin* list:

  *PY@m:n*; *len@offset*; *PY-*$C_k$*@b-offset*, *PY-*$C_k$*@n-offset* or *PY-*$C_k$*@e-offset*

**Statistical Features**. I use pointwise mutual information (PMI) variant (Church and Hanks, 1990) to account for global, corpus-wide information. This measures the difference between the observed probability of an event (*i.e.*, several characters combined as an informal word) and its expectation, based on the probabilities of the individual events (*i.e.*, the

[3] `http://www.bckelk.uklinux.net/menu.html`

probability of the individual characters occurring in the corpus). Compared with other standard association measures such as MI, PMI tends to assign rare events higher scores. This makes it a useful signal for IWR, as it is sensitive to informal words which often have low frequency. However, the word frequency alone is not reliable enough to distinguish informal words from uncommon but formal words.

In response to this challenge in differentiating linguistic registers, I compute two different PMI scores for character-based bigrams from two large corpora representing news and microblogs as features. I also use the difference between the two PMI scores as a differential feature. In addition, I also convert all the character-based bigrams into *pinyin*-based bigrams (ignoring tones[4]) and compute the *pinyin*-level PMI in the same way. These features capture inconsistent use of the bigram across the two domains, which assists to distinguish informal words. Note that I eschew smoothing in the computation of PMI, as it is important to capture the inconsistent character bigrams usage between the two domains. For example, the word *"rp"* appears in the microblog domain, but not in news. If smoothing is conducted, the character bigram *"rp"* will be given a non-zero probability in both domains, not reflective of actual use. For each character $C_i$, I incorporate the PMI of the character bigrams as follows:

- (*) If $C_kC_{k+1}\,(i-4<k<i+4)$ is not a Chinese word recorded in dictionaries:

  *CPMI-N@k+i*; *CPMI-M@k+i*; *CDiff@k+i*; *PYPMI-N@k+i*; *PYPMI-M@k+i*; *PYDiff@k+i*

[4] The informal word may have the same *pinyin* transcription as its formal counterpart without considering the differences in tones.

## 3.2 Experiment

In this section, I discuss *Dataset A*, baseline systems and experiments results in detail in the following.

### 3.2.1 Data Preparation

Utilizing the previous work in Chinese social media archiving, I obtain Chinese microblog posts from the public timeline of Sina Weibo[5] via PrEV (Cui et al., 2012). Sina Weibo is the largest microblogging in China, where over 100 million Chinese microblog posts are posted daily (Cao, 2012), likely the largest public source of informal and daily Chinese language use. *Dataset A* has a total of 6,678,021 messages, covering two months from June to July of 2011. I employed *Zhubajie*[6], a popular crowdsourcing platform in mainland China (Wang et al., 2010) to obtain informal word annotations. In total, I spent US$110 on assembling a subset of $5,500$ posts ($12,446$ sentences) in which $1,658$ unique informal words are annotated within five weeks via Zhubajie. Each post was annotated by three annotators with moderate (0.57) inter-annotator agreement measured by Fleiss' $\kappa$ (Joseph, 1971), and conflicts were resolved by majority voting. Annotations were completed after a five-week span and are publicly available[7] for comparative study.

I divided the annotated corpus, taking $4,000$ posts for training, and the remainder ($1,500$) for testing. Through inspection, it is notable that 79.8% of the informal words annotated in the testing set are not covered by the training set. I also follow my earlier work's conventions Wang et al. (2012a), applying rulesets to preprocess the corpus' *URLs*, *emoti-*

[5] `http://open.weibo.com`

[6] `http://www.zhubajie.com`

[7] `http://wing.comp.nus.edu.sg/portal/downloads.html`

*cons*, *"@usernames"* and *Hashtags* as pre-segmented words, before input to CWS and IWR. For the CWS task, the first author manually labeled the same corpus following the segmentation guidelines published with the *SIGHAN-5*[8] MSR dataset.

### 3.2.2 Baseline Systems

I implemented several baseline systems to compare with proposed FCRF joint inference method.

**Existing Systems**. I re-implemented Wong and Xia (2008)'s extended Support Vector Machine (SVM) based microtext IWR system to compare with my method. Their system only does IWR, using the CWS and POS tagging output of the ICTCLAS segmenter (Zhang et al., 2003) as input. To compare the joint inference versus other learning models, I also employed a decision tree (DT) learner, equipped with the same feature set as the FCRF. Both the SVM and DT models are provided by the Weka3 (Hall et al., 2009) toolkit, using its default configuration.

To evaluate CWS performance, I compare with two recent segmenters. Sun and Xu (2011)'s work achieves the state-of-the-art performance and is publicly available. They employ a LCRF taking as input both lexical and statistical features derived from unlabeled data. As a second baseline, I also evaluate against a widely-used, commercially-available alternative, the recently released 2011 ICTCLAS segmenter[9].

**Two-stage Sequential Systems**. To benchmark the improvement that the factorial CRF model has by doing the two tasks jointly, I compare with a LCRF solution that chains these two tasks together. For completeness, I test pipelining in both directions – CWS feeding features for IWR

[8] `http://www.sighan.org`

[9] `http://www.ictclas.org/index.html`

($LCRF_{cws} \succ LCRF_{iwr}$), and the reverse ($LCRF_{iwr} \succ LCRF_{cws}$). I modify the open-source Mallet GRMM package Sutton (2006) to implement both this sequential LCRF model and the proposed FCRF model. Both models take the whole feature set described in Section 3.1.5.

**Upper Bound Systems**. To measure the upper-bound achievable with perfect support from the complementary task, I also provided gold standard labels of one task (e.g., IWR) as an input feature to the other task (e.g., CWS). These systems (hereafter denoted as LCRF≻LCRF-UB and FCRF-UB) are meant for reference only, as they have access to answers for the opposing tasks.

**Adapted SVM for Joint Classification**. For completeness, I also compared this work against the standard SVM classification model that performs both tasks by predicting the cross-product of the CWS and IWR individual classes (in total, 8 classes). I train the SVM classifier on the same set of features as the FCRF, by providing the cross-product of two layer labels as gold labels. This system (hereafter denoted as SVM-JC) was implemented using the LibSVM package (Chang and Lin, 2011).

### 3.2.3 Evaluation Metrics

I use the standard metrics of precision, recall and $F_1$ for the IWR task. Only words that exactly match the manually-annotated labels are considered correct. For example given the sentence “怎么**介么**好吃呢” (“怎么**这么**好吃呢”; “How delicious it is”), if the IWR component identifies “介么” as an informal word, it will be considered correct, whereas both “介么好” and “介” are deemed incorrect. For CWS evaluation, I employ the conventional scoring script provided in *SIGHAN-5*, which also provides out-of-vocabulary recall (OOVR).

To determine statistical significance of the improvements, I also compute paired, one-tailed $t$ tests. As pointed out by Yeh (2000), the randomization method is more reliable in measuring the significance of $F_1$ through handling non-linear functions of random variables. Thus I employ Padó (2006)'s implementation of randomization algorithm to measure the significance of $F_1$.

### 3.2.4 Experimental Results

The goal of these experiments is to answer the following research questions:

RQ1 Do the two tasks of CWS and IWR benefit from each other?

RQ2 Is jointly modeling both tasks more efficient than conducting each task separately or sequentially?

RQ3 What is the upper bound improvement that can be achieved with perfect support from the opposing task?

RQ4 Are the features designed for the joint inference method effective?

RQ5 Is there a significant difference between the performance of the joint inference of a cross-product SVM and the proposed FCRF?

**CWS Performance.**

In general, the FCRF yields the best performance among all systems (top portion of Table 3.1), answering RQ1. Given microblog posts as test data, the $\mathrm{F}_1$ of ICTCLAS drops from 0.985[10] to 0.698, clearly showing the difficulty of processing microtext. The sequential LCRF model and FCRF model both outperform the baselines, which means with the novel features

[10] Self-declared segmentation accuracy on formal text.http://www.ictclas.org/

shared by the two tasks, CWS benefits significantly from the results of IWR. Hence this segmenter outperforms the existing segmenters by tackling one of the bottlenecks of recognizing informal words in Chinese microtext.

Table 3.1: Performance comparison on the CWS task. The two bottom-most rows show upper bound performance. '‡'('*') in the top four lines indicates statistical significance at $p < 0.001$ (0.05) when compared with the previous row. Symbols in the bottom two lines indicate significant difference between upper bound systems and their corresponding counterparts.

| | **Pre** | **Rec** | **F$_1$** | **OOVR** |
|---|---|---|---|---|
| **ICTCLAS Zhang et al. (2003)** | 0.640 | 0.767 | 0.698 | 0.551 |
| **LCRF Sun and Xu (2011)** | 0.661$^‡$ | 0.691$^‡$ | 0.675 | 0.572$^‡$ |
| **LCRF$_{iwr}$≻LCRF$_{cws}$** | 0.741$^‡$ | 0.775$^‡$ | 0.758$^*$ | 0.607$^*$ |
| **FCRF** | **0.757$^‡$** | **0.801$^‡$** | **0.778$^*$** | **0.633$^*$** |
| **LCRF$_{iwr}$≻LCRF$_{cws}$-UB** | 0.807$^‡$ | 0.815$^‡$ | 0.811$^*$ | 0.731$^‡$ |
| **FCRF-UB** | 0.820$^‡$ | 0.833$^‡$ | 0.826$^*$ | 0.758$^‡$ |

To illustrate, the sequence "...有木有人..." ("...有没有人..."; "...is there anyone..."), is correctly labeled as **BIES** by the FCRF model but mislabeled by baseline systems as **SSBE**. This is likely due to the ignorance of the informal word "有木有", leading baseline systems to keep the formal word "有人" ("someone") as a segment.

More importantly, by jointly optimizing the probabilities of labels on both layers, the FCRF model slightly but significantly improves over the sequential LCRF method, answering RQ2. Thus I conclude that jointly modeling both tasks is more effective than performing the tasks sequentially.

For RQ3, the last two rows present the upper-bound systems that have access to gold standard labels for IWR. Both upper-bound systems statistically outperform their counterparts, indicating that there is still room to improve CWS performance with better IWR as input. This also validates the assumption that CWS can benefit from joint consideration of IWR.

Taking the best previous work as the lower bound (0.69 $F_1$), the FCRF methodology (0.77) makes significant progress towards the upper bound (0.82).

**IWR Performance.**

For RQ1 and RQ2, Table 3.2 compares the performance of my method with the baseline systems on the IWR task. Overall, the FCRF method again outperforms all the baseline systems. It is notable that the CRF based models achieve much higher precision score than baseline systems, which means that the CRF based models can make accurate predictions without enlarging the scope of prospective informal words. Compared with the CRF based models, the SVM and DT both over-predict informal words, incurring a larger precision penalty. Studying this phenomenon more closely, I find it is difficult for the baseline systems to classify segments mixed with formal and informal characters. Taking the microblog "怎么**介么**好吃呢" ("怎么**这么**好吃呢"; "how delicious it is") as an example, without considering the possible word boundaries suggested by the contextual formal words – i.e., "怎么" ("how") and "好吃" ("delicious") – the baselines chunk the informal words (i.e., "介么") together with adjacent characters mistakenly as " 介么好" or, "么介么".

Table 3.2: Performance comparison on the IWR task. '‡' or '*' in the top four rows indicates statistical significance at $p < 0.001$ or $< 0.05$ compared with the previous row. Symbols in the bottom two rows indicate differences between upper bound systems and their counterparts.

| | **Pre** | **Rec** | **$F_1$** |
|---|---|---|---|
| **SVM** | 0.382 | 0.621 | 0.473 |
| **DT** | $0.402^*$ | $0.714^*$ | $0.514^*$ |
| **LCRF$_{cws}$≻LCRF$_{iwr}$** | $0.858^‡$ | $0.591^‡$ | $0.699^*$ |
| **FCRF** | **0.877**$^*$ | **0.655**$^*$ | **0.750**$^*$ |
| **LCRF$_{cws}$≻LCRF$_{iwr}$-UB** | 0.840 | $0.726^*$ | $0.779^*$ |
| **FCRF-UB** | 0.878 | $0.752^*$ | $0.810^*$ |

As indicated by the bold figures in Table 3.2, the FCRF performs slightly better than the sequential LCRF ($p < 0.05$) – a weaker trend when compared with the CWS case. As an example, the sequential LCRF method fails to recognize “爱疯” (“iPhone”) as an informal word in the sentence “我的**爱疯**好玩” (“my iPhone is fun”), where the FCRF succeeds. Inspecting the output, the LCRF segmenter mislabels “爱疯” as **SS**. By jointly considering the probabilities of the two layers, the FCRF model infers better quality segmentation labels, which in turn enhances the FCRF’s capability to recognize the sequence of two characters as an informal word. This is further validated by the significant performance gulf between the upper bound and the basic system shown in the lower half of the table.

For RQ3, interestingly, the difference in performance between the LCRF and FCRF upper-bound systems is not significant. However, these are upper bounds, and I expect on real-world data that CWS performance will not be perfect. As such, I still recommend using the FCRF model, as the joint process is more robust to noisy input from one channel.

**Feature Set Evaluation.**

For RQ4, to evaluate the effectiveness of the newly-introduced feature sets (those marked with “*” in Section 3.1.5), I also test a FCRF ($\mathbf{FCRF}_{-new}$) without the new features. According to Table 3.3, performance drops by a significant amount: 0.088 $F_1$ on CWS and 0.198 $F_1$ on IWR. $\mathbf{FCRF}_{-new}$ makes many mistakes identical to the baselines: segmenting informal words into several single-character words and chunking adjacent characters from informal and formal words together.

Table 3.3: $F_1$ comparison between FCRF and FCRF$_{-new}$. ('*') indicates statistical significance at $p < 0.05$ when compared with the previous row.

| | **CWS** | **IWR** |
|---|---|---|
| **FCRF$_{-new}$** | 0.690 | 0.552 |
| **FCRF** | $0.778^*$ | $0.750^*$ |

### Adapted SVM-JC vs. FCRF.

For RQ5, according to Table 3.4, the SVM trained to predict the cross-product CWS/IWR classification (SVM-JC) performs quite well on its own. Unsurprisingly, it does not outperform the proposed FCRF, which has access to more structural correlation among the CWS and IWR labels. SVM-JC significantly ($p < 0.001$) outperforms the baseline SVM system by 0.151 in the IWR task, which I think is partially explained by its good performance (0.761) on the CWS task. The tendency of the model to overly predict positive results in the individual SVM case is effectively addressed by simultaneously modeling the CWS task, whereas FCRF turns out to be more effective in solving joint inference problem, although in a weaker trend in terms of the statistical significance ($p < 0.05$).

Table 3.4: $F_1$ comparison between SVM, SVM-JC and FCRF. '‡'('*') indicates statistical significance at $p < 0.001$ (0.05) when compared with the previous row.

| | **CWS** | **IWR** |
|---|---|---|
| **SVM** | — | 0.473 |
| **SVM-JC** | 0.741 | $0.624^‡$ |
| **FCRF** | $0.778^*$ | $0.750^*$ |

I conclude that the use of the FCRF model and the addition of new features are both essential for the high performance of this system.

## 3.3 Discussion

It is desirable to understand the causes of errors in the model so that we may better understand its weaknesses. Manually inspecting the errors of the system, I found three major categories of errors which are dissected here.

For IWR, the major source of error, accounting for more than 60% of all errors, is caused by what I term the *partially observed informal word* phenomenon. This refers to informal words containing multiple characters, where some of its components have appeared in the training data as informal words individually. For instance, the single-character informal word, “狠” (“很”; “very”) appears in training multiple times, thus the unseen informal word “狠久” (“很久”; “long time”) is a *partially observed* informal word. In this case, the model incorrectly labels the known, single character “狠” with **IF_S** as an informal word, instead of labeling the unseen sequence “狠久” with correct labels **IF_B IF_E**. Errors then result in both tasks.

This observation motivates the use of the relation between the known informal word and its formal counterpart in order to inform the model to better predict in cases of partial observations. Following the same example, given that “狠” is an informal word, if the model also considers the probability of normalizing “狠” to “很”, while considering the higher probability that the character sequence “很久” could be a formal word, there would be a higher likelihood of correctly predicting the sequence “狠久” as an informal word. So informal word normalization is also an intrinsic component of IWR and CWS, and I believe it is an interesting direction for future work.

Another source of error is a side effect of microtext being extremely short. For example, in the sentence “**肥家！**太累了。。。” (“**回家！**太累了。。。”; “Go home! Exhausted.”), the unseen informal word “肥家” itself forms a short sentence. Although it has a subsequent sentence “太累了。。。” (“Exhausted”) as context, and the two are pragmatically related, (i.e., “I am exhausted! [And as a result,] I want to go home.”), the lexical relationship between the sentences is weak; i.e., “太累了。。。” appears frequently as the context of various sentences, making the context difficult to utilize. These phenomena makes it difficult to recognize “肥家” as an informal word.

A possible solution could factor in proximity, similar to density-based matching, as in Tellex et al. (2003). It is reasonable to assign a higher weight to features related to characters closer to the current target character. In particular, for this example, given the current target character “肥”, we can assign higher weight to features generated from features from the proximal context “肥家”, and lower weight to features extracted from distal contexts.

Another major group of errors come from what I term *freestyle named entities* as exemplified in Table 3.5; i.e., person names in the form of user IDs and nicknames, that have less constraint on form in terms of length, canonical structure (not surnames with given names; as is standard in Chinese names) and may mix alphabetic characters. Most of these belong to the category of *Person* (PER), as defined in CoNLL-2003[11] Named Entity Recognition shared task. Such freestyle entities are often mistakenly recognized as informal words, as they share some of the same stylistic markings, and are not marked by features used to recognize previous Chinese named entity recognition methods (Gao et al., 2005; Zhao and Kit, 2008) that work on news or general domain text. As described in the introduction to

[11] `http://www.cnts.ua.ac.be/conll2003/ner/`

Table 3.5: Sample Chinese freestyle named entities that are usernames.

| **Freestyle Named Entity** | **Explanation** |
|---|---|
| “榴莲雪媚娘” | “榴莲” (“durian”), “雪” (“snow”), “媚娘” (“charming lady”) |
| “ 棉宝” | It is short for the cartoon name “海绵宝宝”. |
| “dj文祥”, “徐pp” | Usernames mixed of Chinese and alphabetic characters |

this thesis, I have recognized this as a challenge in Chinese microtext and tackle it later in Chapter 5.

## 3.4 Summary

There is a close dependency between Chinese word segmentation (CWS) and informal word recognition (IWR). To leverage this, I employ a factorial conditional random field to perform both tasks of CWS and IWR jointly.

I propose novel features including statistical and lexical features that improve the performance of the inference process. I evaluate the method on a manually-constructed data set and compare it with multiple research and industrial baselines that perform CWS and IWR individually or sequentially. The experimental results show the joint inference model yields significantly better $F_1$ for both tasks. For analysis, I also construct upper bound systems to assess the potential maximal improvement, by feeding one task with the gold standard labels from the complementary task. These experiments further verify the necessity and effectiveness of modeling the two tasks jointly, and point to the possibility of even better performance with improved per-task performance.

Analyzing the classes of errors made by the system, I identify a promising future work topic to handle errors arising from *partially observed informal words* – where parts of a multi-character informal word have been observed before. Incorporating informal word normalization into the inference process may help address this important source of error.

# Chapter 4

# Chinese Informal Word Normalization

In the second key area, I explore the linguistic characteristics of existing informal words in Chinese microtext. Informed by prior work and my survey, I find that informal Chinese words largely originate from their formal equivalents by means of one of three key channels. I present a novel method for normalizing Chinese informal words to their original formal equivalents. This chapter describes my method for building a Chinese informal word normalization system. I formalize the task as a classification problem and propose rule-based and statistical features to model the connections between informal word and its origin formal word. I evaluated on a crowdsourced corpus, achieving a normalization precision of 89.5% across all the different channels.

## 4.1 Data Analysis

To bootstrap this work, I analyzed sample Chinese microtext, hoping to gain insight on how informal words relate to their formal counterparts.

Starting from *Dataset A* as introduced in Chapter 3.2, again, I employed the same crowdsourcing platform to obtain third-party (*i.e.*, not by the original author of the microtext, due to likely unavailability) annotations for their normalization, sentiment and motivation for its use (Wang et al., 2010). The coarse-grained sentiment annotations use three categories of "positive", "neutral" and "negative". Motivation is likewise annotated with the seven categories listed in Table 4.1:

Table 4.1: Categories used for motivation annotation, shown with their observed distribution.

| | |
|---|---|
| to avoid (politically) **sensitive** words | 17.8% |
| to be **humorous** | 29.2% |
| to hedge criticism using **euphemisms** | 12.1% |
| to be **terse** | 25.4% |
| to **exaggerate** the post's mood or emotion | 10.5% |
| **others** | 5.0% |

I collected this set of crowdsourced annotations partly for posterity and as such, not all of the annotations collected are pertinent to my thesis. In particular, the collected sentiment annotations were only stored, but not used in any work related to this thesis. I mention it here for completeness.

### 4.1.1 Data Feature Analysis

From my observation of the annotated informal/formal word pairs, I identified three key channels through which the majority of informal words originate, summarized in Table 4.2. Here, the first column describes the channels, giving each channel's observed prevalence as a percentage. Together, they account for about 94% of the instances by which informal words originate. The final "Motivation (%)" column also gives the distributional breakdown of motivations behind each of the channels as annotated by those crowdsourced annotators. I now discuss each channel.

Table 4.2: Classification of Chinese informal words as originating from three primary channels. Pronunciation is indicated with *pinyin* for phonetic substitutions, while characters in bold are linked to the motivation for the informal form.

| Channel (%) | Informal | Formal | Translation | Sentiment | Motivation (%) |
|---|---|---|---|---|---|
| **Phonetic Substitutions** (63) | 河蟹 | 和谐 | harmonious | positive | **sensitive** (28.9) |
| | 鸭梨 | 压力 | pressure | neutral | **humorous** (45.2) |
| | **bs** | 鄙视 | despise | negative | **euphemism** (18.7) |
| | 足已 | 足以 | sufficient to | neutral | **others** (7.2) |
| **Abbreviation** (19) | 桌游 | 桌面␣游戏 | board game | neutral | **terse** |
| | 剧透 | 剧情␣透露 | tell spoilers | neutral | **terse** (100) |
| **Paraphrase** (12) | 给力 | 很棒 | awesome | positive | **exaggerate** |
| | 暴汗 | 非常␣尴尬 | very awkward | negative | **terse** (27.3) |
| | 卖萌 | 可爱 | cute | positive | **others** (6.4) |

**Phonetic Substitutions** form the most well-known channel where the resultant informal words are pronounced similar to their formal counterparts. It is also the channel responsible for most informal word derivation. It has been reported to account for 49.1% (Li and Yarowsky, 2008a) of informal words in the Web domain and for 99% in Chinese chats (Xia et al., 2006). In my study of the microtext domain, it is found to be responsible for 63% (Table 4.2). As highlighted in bold in the table, normalization in this channel is realized by a **character–character** *pinyin* mapping. An interesting special case occurs when the Chinese characters are substituted for Latin alphabets, where the alphabets form a *pinyin* acronym. In these cases, each letter maps to a *pinyin* initial (e.g., "bs" → 'b"+ "s" → "bi" + "shi" ("鄙视(**bi shi**)"; "to despise")), each of which maps to a single Chinese character. As such, I view this special case as also following the character–character mapping.

I found that phonetic substitutions are motivated by different intents. Slightly over half of the words are used to be humorous. This resonates well with the informal context of many microtexts, such that authors take advantage of expressing their humor through lexical choice. Another large group (28.9%) of informal words are variations of *politically sensitive words* (*e.g.*, the names of politicians, religious movements and events), whose formal counterparts are often forbidden and censored by search engines or Chinese government officials. Netizens often create such phonetically equivalent or close variations to express themselves and communicate with others on such issues. An additional 18.7% of such word pairs are used euphemistically to avoid the usage of their harsher, formal equivalents. The remaining substitutions are explainable as typographical errors, transliterations, among other sources.

The **Abbreviation** channel contains informal words that are shortenings of formal words. Normalizing these informal words is equivalent to expanding short forms to corresponding full forms. As suggested by Chang and Teng (2006), Chinese abbreviation expansion can be modeled as **character–word** mapping. The statistics in Table 4.2 suggest 19% of informal words come from this channel, and are used to save space and to make communication efficient, especially given the format and length limitations in microtext.

**Paraphrases** mark informal words that are created by a mixture of paraphrasing, abbreviating and combining existing formal words. I observe that the informal manifestation usually do not retain any of the original characters in their formal equivalents, but still retain the same meaning as a single formal word, or two meanings combined from two formal words. These words are created to enhance emotional response in an exaggerated (66.3%) and/or terse (27.3%) manner. For example in Table 4.2, "给力" as a whole comes from the paraphrase of the single formal word "很棒", sharing the meaning of "awesome". Another example, "暴汗" ["very embarrassed"] originates from two sources: "暴" meaning "十分" ["very"] and "汗" meaning "尴尬" ["embarrassed"]. From these observations, I feel that both **character–word** and **word–word** mappings may adequately model the normalization process for this channel.

## 4.2 Methodology

Drawing on my observations, I propose a two-step generation—classification model for informal word normalization. I first *generate* potential formal candidates for an input informal word by combing through the Google 1T corpus. This step is fast and generates a large, prospective set of candidates which are input to a second, subsequent classification. The subsequent classification is a binary yes/no classifier that takes both rule-based and statistical features derived from three identified major channels to identify valid formal candidates.

Note that an informal word $O$ (here, $O$ for observation), even when used in a specific, windowed context $C(O)$, may have several different equivalent normalizations $T$ (here, $T$ for target). This occurs in the abbreviation ("桌游" ["board game"] as "桌面游戏" ["board game"]) and paraphrase ("给力" ["awesome"] "很棒" ["great"] or "很好" ["very good"] or "厉害" ["awesome"]) channels, where synonymous formal words are equivalent. In the case where an informal word is explainable as a phonetic substitution, only one formal form is viable. The proposed classification model caters for these multiple explanations.

Figure 4.1 illustrates the framework of the proposed approach. Given an input Chinese microblog post, I first segment the sentences into words and recognize informal words leveraging the approach proposed in Wang and Kan (2013). For each recognized informal word $O$, I search the Chinese portion of the Google Web 1T corpus using lexical patterns, obtaining $n$ potential formal (normalized) candidates. Taking the informal word $O$, its occurrence context $C(O)$, and the formal candidate $T$ together, I gener-

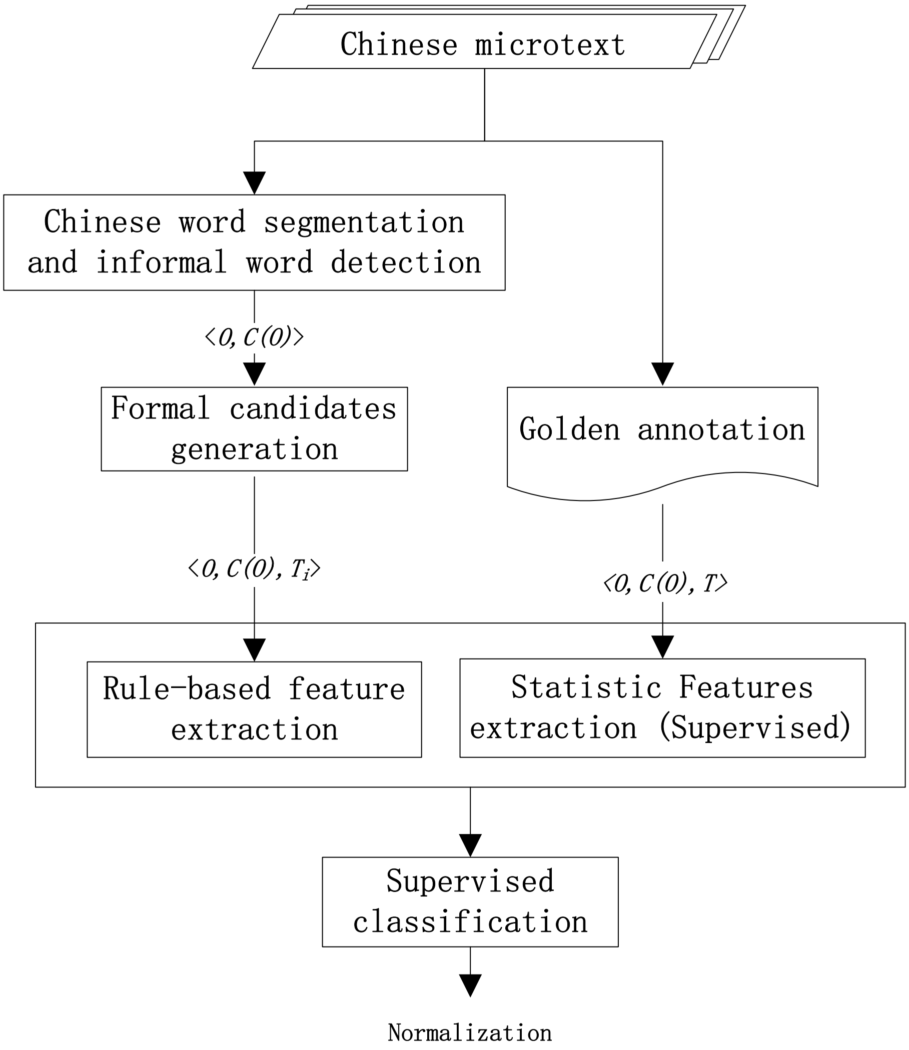

Chinese microtext
Chinese word segmentation and informal word detection
<O, C(O)>
Formal candidates generation
Golden annotation
<O, C(O), Ti>
<O, C(O), T>
Rule-based feature extraction
Statistic Features extraction (Supervised)
Supervised classification
Normalization

Figure 4.1: The framework consists of the two steps of informal word recognition and normalization. Normalization breaks down to its component steps of candidate generation and classification.

ate feature vectors for each three-tuple, i.e., $< O, C(O), T >$[1], consisting of both rule-based and statistical features. These features are used in a supervised binary classifier to render the final yes (informal–formal pair) or no (not an appropriate formal word explanation for the given informal word) decision.

[1] For notational convenience, the informal word context $C(O)$ is defined as $W_{-i}...O\ ...W_i$; here, $i$ refers to the index of the word with respect to $O$, which is set in this work to 3.

### 4.2.1 Pre-Processing

As an initial step, it is possible to recognize informal words and segment the Chinese words through the method introduced in Chapter 3. However, as the focus in this work is on the normalization task, I use the manually-annotated gold standard informal words ($O$) and their formal equivalents ($T$) provided in the annotated dataset. To derive the informal words' context $C(O)$, I use the automatically-acquired output of the preprocessing FCRF, although noisy and a source of error.

### 4.2.2 Formal Candidate Generation

Given the two-tuple $< O, C(O) >$ generated from pre-processing, I produce a set of hypotheses $|T|$ which are formal candidates corresponding to $O$. I use two assumptions to guide the selection of prospective formal equivalents of $O$. I first discuss Assumption 1 (as [A1]):

[A1] The informal word and its formal equivalents share similar contextual collocations.

To implement [A1], I define five regular expression patterns to search the Chinese Web 1T corpus, listed in Table 4.3. All entries that match at least one of the five rules are collected as formal candidates. Specifically, $W_*$ refers to the word in context $C(O)$. $T$ denotes any Chinese candidate word, and $\hat{T}$ a word sharing at least one character in common with the informal word $O$.

Table 4.3: Lexical patterns for candidate generation.

| | | |
|---|---|---|
| $W_{-1}\ T\ W_1$ | $W_{-2}\ W_{-1}\ T$ | $T\ W_1\ W_2$ |
| $W_{-1}\ \hat{T}$ | | $\hat{T}\ W_1$ |

My assumption is similar to the notion used for paraphrasing: that the informal version can be substituted for its formal equivalent(s), such that the original sentence's semantics is preserved in the new sentence. For example, in the phrase "建设␣河蟹␣社会" ["build the harmonious society"], the informal word "河蟹" ["harmonious"] is exactly equivalent to its formal equivalent "和谐" ["harmonious"], as the resulting phrase "建设␣和谐␣社会" ["build the harmonious society"]) carries exactly the same semantics. This is inferrable when both the informal word $O$ and the candidate share the same contextual collocations of "建设" ["build"] and "社会" ["society"].

As the Web 1T corpus consists of $n$-grams taken from approximately one trillion words indexed from Chinese web pages, queries for each informal word $O$ can return long result lists of up to 20,000 candidates. To filter noise from the resulting candidates, I adopt Assumption 2 [A2]:

[A2] Both the original informal word in its context – as well as the substitued formal word within the same context – are frequent in the general domain.

I operationalize this by constraining the prospective normalization candidates to be within the top 1,000 candidates ranked by the trigram probability ($P(W_{-1}\ T\ W_1)$). This probability is calculated by the BerkeleyLM (Pauls and Klein, 2011) trained over Google Web 1T corpus. Note that this constraint makes the method more efficient over a brute-force approach, in exchange for loss in recall. However this trade-off is fair: by retaining the top 1000 candidates, I observed the loss rate of gold standard answers in each of the channels is 14%, 15%, and 17% for phonetic substitution, abbreviation and paraphrase, respectively. This is in comparison with the final loss rate of over 70% reported by Li and Yarowsky (2008a).

Given the annotations, the three-tuples ($< O, C(O), T >$) generated from the resulting list of candidates are labeled as **Y** (**N**) as positive (negative) instances. As there are a much larger number of negative than positive instances for each $O$, this results in data skew.

### 4.2.3 Feature Extraction for Classification

For the classification step, I calculate both rule-based and statistical features for supervised machine learning. I leverage previous observations to engineer features specific to a particular channel. I describe both classes of features, listing its type (*b*inary or *c*ontinuous) and which channel it models (*ph*onetic substitution, *ab*breviation,*pa*raphrase, or *all*), as a two tuple. I describe each rule accompanied by an example, showing *pinyin* and tones, where appropriate.

**Rule-based Features (5 features)**

- $O$ contains valid *pinyin* script $< b, ph >$
  e.g., “ 冻shi 了” (“ 冻死si3了”;“too cold”)
- $O$ contains digits $< b, ph >$
  e.g., “ v5” (“威wei1武wu3”;“mighty”)
- $O$ is a potential *pinyin* acronym $< b, ph >$
  e.g., “bs” (“鄙bi3视shi4”;“despise”)
- $T$ contains characters in $O$? $< b, ph >$
  e.g., “ 桌游” (“ 桌面游戏”;“board games”)
- The percentage of characters common between $O$ and $T$ $< c, all >$

**Statistical Features (7 features)**

I describe these features in more detail, as they form a key contribution in this work. Note that the statistical features that leverage information from both informal and formal domains are derived via maximum likelihood estimation on the appropriate training data.

*Pinyin* **Similarity** $< c, ph >$. Although Levenshtein distance ($LD$; employed in (Li and Yarowsky, 2008a)) is a low cost metric to measure string similarity, it has its drawbacks when applied to *pinyin* similarity. As an example, the informal word " 淫yin2 才cai2 " ["talent"] is normalized to "人ren2 才cai2" ["talent"] , meaning "talent". This suggests that $PYSim(yin, ren)$ should be high, as they compose an informal-formal pair. However this is in contrast to evidence given by $LD$, as $LD(yin, ren)$ is large (especially compared with the $LD(yin, yi)$, in which "yi" is a representative *pinyin* string that has an edit distance with "yin" of just 1). For the manual annotation method, it is difficult for annotators to accurately weigh the similarities for all pronunciation pairs, since it is weighted arbitrarily. And the labor of manually tuning weights may be unnecessary, given annotated informal–formal pairs.

To tackle these drawbacks, I propose to fully utilize the gold standard annotation (*i.e.*, informal–formal pairs applicable to the Phonetic Substitution channel) and to empirically estimate the *pinyin* similarity from the corpus in a supervised manner. In this method, *pinyin* similarity is formulated as:

$$PYSim(T|O) = \prod PYSim(t_i|o_i) \tag{4.1}$$

$$\begin{aligned} & PYSim(t_i|o_i) = PYSim(py(t_i)|py(o_i))) \\ & = \mu P(py(t_i)|py(o_i)) + \lambda P(ini(t_i)|py(o_i)) \\ & + \eta P(fin(t_i)|py(o_i)) \end{aligned} \tag{4.2}$$

Here, the $ti$ $(o_i)$ stands for the $i$th character in word $T$ $(O)$. Let the function $py(x)$ return the *pinyin* string of a character and functions $ini(x)$ $(fin(x))$ return *initial* (*final*) of a *pinyin* string $x$. I use linear interpolation algorithm for smoothing, with $\mu$, $\lambda$ and $\eta$ as weights summing to unity. Then, $P(py(t_i)|py(o_i))$, $P(ini(t_i)|py(o_i))$ and $P(fin(t_i)|py(o_i))$ are estimated using maximum likelihood estimation over the training set.

**Lexicon and Semantic Similarity** $< c, ab+pa >$. For the remaining two channels, I extend the source channel model (SCM) (Brown et al., 1990) to estimate the character mapping probability. In my case, SCM aims to find the formal string $T$ that the given input $O$ is most likely normalized to.

$$\hat{T} = \arg\underset{T}{max} P(T|O) = \arg\underset{T}{max} P(O|T)P(T) \tag{4.3}$$

As discussed in Section 3, for both the two channels I use interpolation to model character–word mappings. Assuming the character–word mapping events are independent, I obtain:

$$P(O|T) = \prod P(o_i|t_i) \tag{4.4}$$

where $o_i$ $(t_i)$ refers to $i$th character of $O$ $(T)$. However, this SCM model suffers serious data sparsity problems, when the annotated microtext corpus is small (as in my case). To further address the sparsity, I extend the source channel model by inserting part-of-speech mapping models into Equation 4.4.

$$P(O|T) = \prod P'(o_i|t_i) \tag{4.5}$$

$$P'(o_i|t_i) \quad = \alpha P(o_i|t_i) + \beta P(o_i|pos(t_i), pos(o_i)) \tag{4.6}$$

Here, let the function $pos(x)$ return the part-of-speech (POS) tag of $x$[2]. Both $P(o_i|t_i)$ and $P(o_i|pos(t_i), pos(o_i))$ are then estimated using maximum likelihood estimation over the annotated corpus. In parallel with the *pinyin* similarity estimation, $\alpha$ and $\beta$ are weights for the interpolation, summing to unity.

I give the intuition for the formulation. $P(o_i|t_i)$ measures the probability of using character $o_i$ to substitute for the given word $t_i$. $P(o_i|pos(t_i), pos(o_i))$ measures the probability of using character $o_i$ as the substitution of any word $t_i$, given the POS tag is mapped from $pos(t_i)$ to $pos(o_i)$. Finally, given the limited availability of gold standard annotations, it is optional to use formal domain synonym lexica to improve the model's estimation lexical and semantic similarity.

**N-gram Probabilities** $5\times < c, all >$. I generate new sentences by substituting informal words with candidate formal words. The probabilities of the generated trigrams and bigrams (within a window size of 3) are computed with BerkeleyLM, trained on the Web1T corpus. The features capture how likely the candidate word is used in the informal domain. The five features are:

- Trigram probabilities: $P(W_{-2}W_{-1}T)$; $P(W_{-1}T\ W_1)$;$P(T\ W_1W_2)$

- Bigram probabilities: $P(W_{-1}\ T)$; $P(T\ W_1)$

[2] Implemented in my system by the FudanNLP toolkit `https://code.google.com/p/fudannlp/` Qiu et al. (2013)

## 4.3 Experiments

In my architecture, the candidate generation procedure is unsupervised. The part that does need tuning is the final, supervised classifier that renders the binary decision on each 3-tuple, as to whether the *O*–*T* pair is a match, so for this task I select the best classifier among three learners. The statistics reported by Li and Yarowsky (2008a) is then used as a baseline$^*$ performance. I mark this with an asterisk to indicate that the comparison is just for reference, where the performance figures are taken directly from their published work, as I did not re-implement their method nor execute it on the contemporary data.

As a second analysis point, I compare the system – with and without features derived from synonym lexica – to assess how well this method adapts from formal corpora. Finally I show that this method is also effective to acquire synonyms for the formal domain (formal/formal pairs, in contrast to my task's informal/formal pairs).

### 4.3.1 Data Preparation

I collected 1036 unique informal/formal word pairs with their informal contexts were collected from the annotated corpus. As my contribution is in the methodology and not the learner applied, any supervised classifier suffices. I test logistic regression (LR), support vector machine (SVM) and decision tree (DT) learning models, provided by WEKA3 (Hall et al., 2009). To acquire formal domain synonyms, I also (optionally) employed the Cilin[3] and TYCDict[4] lexica.

[3] `http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=162`

[4] `http://www.datatang.com/data/29207/`

### 4.3.2 Results

I adopt the standard metrics of precision, recall and $F_1$ for the evaluation, focusing on the the positive (correctly matched as informal/formal pair) **Y** class.

#### Classifier Choice

Table 4.4 presents the evaluation results over different classifiers. In this first experiment, data from all the channels are merged together and the result reported is the outcome of 5-fold cross validation. Lexicon similarity features are derived only from the relevant training corpus, for each fold, to ensure correct results without peeking. As the DT classifier performs best, only DT results are reported for subsequent experiments.

Table 4.4: Performance comparison using different classifiers.

| **Classifier** | **Pre** | **Rec** | $F_1$ |
|---|---|---|---|
| **SVM** | 0.646 | 0.273 | 0.383 |
| **LR** | .0567 | 0.340 | 0.430 |
| **DT (C4.5)** | 0.886 | 0.443 | 0.590 |

#### Baseline* Comparison

To make a direct comparison with the baseline*, I perform cross-fold validation using data each of three channels separately. Since Li and Yarowsky (2008a) formalized the task as a ranking problem, this work shows the reported Top1 and Top10 precision in Table 4.5[5].

This model achieves high precision for each channel, compared with the baseline* performance. From Table 4.5, I observe that normalizing words due to Phonetic Substitution is relatively easy as compared to the

[5] Due to the difference in classification scheme, I re-computed the reported value, given the classification.

other two channels. That is because given the fixed vocabulary of standard Chinese *pinyin*, the *pinyin* similarity measured from the corpus is much more stable than the estimated lexicon or semantic similarity. The low recall for the Paraphrase channel suggests the difficulty of inferring the semantic similarity between word pairs.

Table 4.5: Performance, analyzed per channel. "—" indicate no comparable prior reported results.

| **Channel** | **System** | **Pre** | **Rec** | $F_1$ |
|---|---|---|---|---|
| **Phonetic Substitution** | **OurDT** | .956 | .822 | .883 |
| | **LY Top1** | .754 | — | — |
| | **LY Top10** | .906 | — | — |
| **Abbreviation** | **OurDT** | .807 | .665 | .729 |
| | **LY Top1** | .118 | — | — |
| | **LY Top10** | .412 | — | — |
| **Paraphrase** | **OurDT** | .754 | .331 | .460 |
| | **LY Top1** | — | — | — |
| | **LY Top10** | — | — | — |

### Final Loss Rate

It is notable that there is a tradeoff between the data scale and performance. By keeping the Top 1000 candidates, I observed an 18.8% overall loss of correct formal candidates (breaking down as 14.9% for Phonetic Substitutions, 22.8% for Abbreviations and 31.8% for Paraphrases). Based on this statistics, the final loss rate is 64.1%. By comparison, Li and Yarowsky (2008a)'s seed bootstrapped method's self-stated loss rate is around 70%.

### Channel Knowledge and Use of Formal Synonym Dictionaries

In the real-world, it is valuable to infer the channel an informal word originates from. To assess how well the system does without channel knowledge, I merged the separate channel datasets together and train a single classifier.

To investigate the impact of the formal synonym lexica, two configurations – with and without features derived from synonym lexica – were also tested. To upper bound achievable performance, I trained an oracular model with the correct channel as an input feature. In the results presented in Table 4.6, I find that the introduction of the features from the formal synonym lexica enhances performance (especially for recall) of the basic feature set. As upper-bound performance is still significantly higher, future work may aim to improve performance by first predicting the originating channel.

Table 4.6: Performance over different feature sets. "w" ("w/o") refers to the model trained with (without) features from formal synonym dictionaries. "channel" refers to the model trained with the correct channel given as an input feature.

| **Feature set** | **Pre** | **Rec** | $F_1$ |
|---|---|---|---|
| **w/o** | 0.886 | 0.443 | 0.590 |
| **w** | **0.895** | **0.583** | **0.706** |
| **w + channel** | 0.915 | 0.638 | 0.752 |

## 4.4 Summary

Based on the observations from a crowdsourced annotated corpus of informal Chinese words, I perform a systematic analysis about how informal words originate. There are three main channels – phonetic substitution, abbreviation and paraphrase – that are responsible for informal creation, and that the motivation for their creation varies by channel.

To operationalize informal word normalization, I propose a two-stage candidate generation-classification method. The results obtained are promising, improving over the current state of the art with respect to both $F_1$ and loss rate. In the detailed analysis, I find that knowledge of the origin channel can still improve performance and is a possible field for future work.

# Chapter 5

# Named Entity Recognition on Microtext

Due to the nature of microtext, I have shown at the outset of this thesis that Named Entities (NEs) commonly occur as part of everyday social communication and gossip. As the recognition of people, places and things are crucial to everyday discourse, I have identified Named Entity Recognition (NER) as a third, critical area of processing that enables downstream analytics. However, NEs in such informal text, exhibit a wider variety than conventional named entities (*i.e.*, persons' name, location and organization) – one must extend the definition of NEs in the informal domain to encompass *Titles* (*i.e.*, titles of the books, movies, TV Shows, games and so on) and *Products* (*i.e.*, general objects offered to a market), which are much more frequently used in microtext. A main difference from the news domain is the effort to deal with ambiguity and its time-sensitive nature. Another key problem is the lack of training data in microblog domain. This can be tackled through crawling the data labeled by punctuations automatically. In this chapter of the thesis, I make my final contribution towards informal Chinese language processing, and propose efficient methods to address this ambiguity, and exploit its time sensitive nature to further improve the performance and reduce the size of training model.

## 5.1 Methodology

To gain insights on the characteristics of *Title* and *Product* NE occurrences, I studied the annotated microblog posts in *Dataset B* introduced in Chapter 1. The observations informed me on strategies that I implemented to address the below issues. Figure 5.1 illustrates an example microblog post graphically, with two named entities (“青岛啤酒” [“QingDao beer”] and “啦啦宝贝大PK” [“cheerleading PK”]).

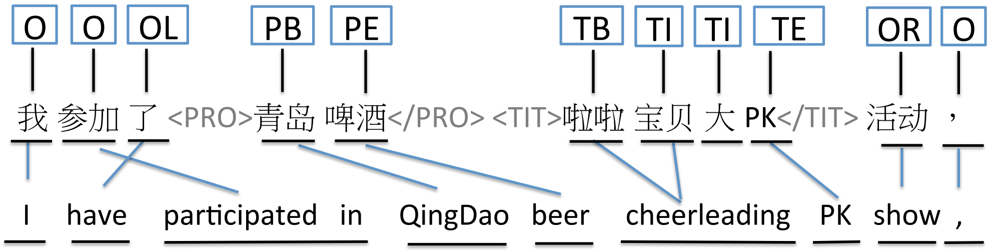

O
O
OL
PB
PE
TB
TI
TI
TE
OR
O
我 参加 了 <PRO>青岛 啤酒</PRO> <TIT>啦啦 宝贝 大 PK</TIT> 活动 ，
I have participated in QingDao beer cheerleading PK show ,

Figure 5.1: A Chinese microtext (bottom layer) with annotations for NER. The bottom line gives the aligned English translation. “TIT”,“PRO” are the annotations of *Title* and *Production*.

### 5.1.1 Issues with Ambiguity

I observe that the specific characters, words and alphabetic letters that are the components of such NEs are also commonly-used with their original semantics (i.e., as non-NEs) in both informal and formal text domain. As a result, a standard supervised strategy employing a training corpus, is unlikely to hypotheses these words (*e.g.*, “啦啦” [“LaLa”], “ 宝贝” [“baby”] and “大” [“big”]) as elements of a NE, unless the NE (*e.g.*, “啦啦宝贝大PK” [“cheerleading PK”]) as a whole has been previously labeled in the corpus. Moreover, unlike the NER in formal domain, which can benefit from multiple lexica (*e.g.*, surname list, location list and a list of common suffixes), the NEs in microtext exhibit a diverse vocabulary but few lexical indicators. This increases the sparsity of lexical features.

Another useful observation is on context: the boundary words directly before and after NEs. As shown in Figure 5.1, "了" ["past tense marker"], the Chinese particle expressing completion is a implicit hint that the following "青岛啤酒" ["Qingdao beer"] is a *Product*. Similarly, the direct object "活动" ["show"] implicitly suggests "啦啦宝贝大PK" ["cheerleading PK"] to be a *Title* of show. Although these indicators are not foolproof, they are strong indicators of NEs and can be utilized in a straightforward manner, making their presence worth leveraging. With this in mind, I propose to employ LCRF model to label NEs together with these boundary words modeled as specifically-designed features. The experiment in Section 5.2 further validates such boundary words' importance through the performance achieved.

### 5.1.2 Issues with Acquiring Training Data

While *Title*s in Chinese formal text are strictly (i.e., grammatically) required to be indicated by paired guillemets (*i.e.*,"《》" ["double angle quotation mark"]), this rule does not apply to informal microtext. Based on my statistical analysis on *Dataset B*, 92.5% of titles that have been labeled do not feature guillemets as indicators; however, crucially, the remaining portion does exhibit them. This offers the opportunity to automatically obtain labeled microtext *Title* instances by using paired guillemets to delimit positive training examples.

To acquire automatically labeled annotations for *Product*, I leveraged manual resources for bootstrapping. I use the entries defined in the Chinese encyclopedia Baidu Baike [1] under their *Product* category as keywords, which are further used to query the Sina Weibo corpus. For each month between February and July of 2011, I randomly collected 5000 posts for use as part of a training corpus. I selected an 1000 additional posts from July 2011 as the evaluation data set, and annotated these. Hereafter I use "*<Label>_<Month>*" (*e.g.*, *TIT_Feb* and *PRO_Jul*) to denote the data sets.

Even with this solution in acquiring training data automatically, such an automated annotation method results in a skewed corpus, which contains mostly positive training instances. To further balance the training corpus, negative instances – especially those containing the potential NE component words or those with boundary words – are needed to balance out the learned model. To address this, I propose the following steps to select negative instances in my approach:

1. Replace all occurrences of *Title* and *Product* with a special "NE" token in the training corpus,

2. Compute the pointwise mutual information (PMI) between boundary words and the special NE token,

3. Construct Queries: Select all the words from NEs together with the boundary words having PMI higher than a threshold $\theta$ as queries,

4. Search instances from the whole corpus using these queries, and

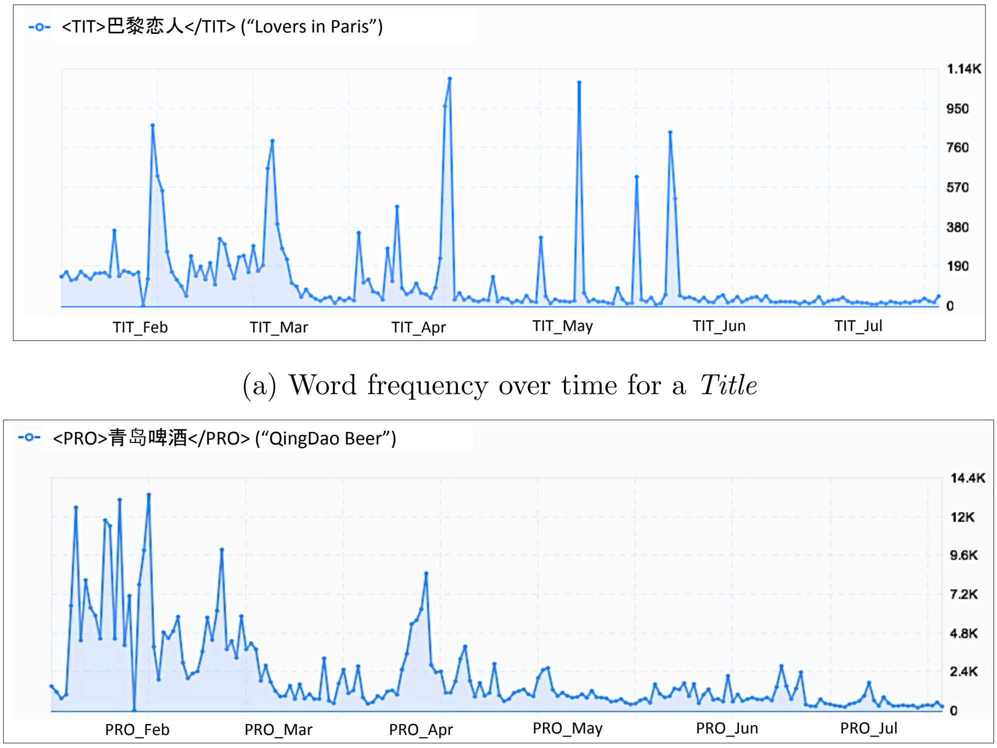
5. Select the instances without any manually-tagged NEs as negative instances.

[1] `www.baike.baidu.com`

The key idea behind these steps is to lessen bias through selectively annotating those negative instances from the microblog posts that contain boundary words or partial NE component words, but not NE tokens. 2000 negative instances are selected and annotated in total.

### 5.1.3 Time Sensitivity

As hot topics in social media, *Title* and *Product* wax and wane, tied with the trends popular events. Figure 5.2 demonstrates the word frequency distribution across six months time period for two sample NEs. In the figure, one sees that the frequent peaks always last for about one week, and the time spans between two adjacent peaks last for about one month. These observations tend to hold across many examined NEs; the figure is representative of temporal distribution of NEs in microtext.



(a) Word frequency over time for a *Title*

(b) Word frequency over time for a *Product*

Figure 5.2: Statistics of two sample NEs.

A key insight from analyzing such NEs over the corpus, is that the intervals between two adjacent frequency peaks are always less than one month. This means if we observe a frequency peak for a particular NE in current month, it is likely to have another frequency peak in the next month. Besides, the frequency always drops to a steady, low-level state after multiple continuous frequency peaks. In other words, if no frequency peak is observed in current month, the chance of having a frequency peak in the next month is relatively low. This implies that the recognition of NEs frequently mentioned in month **T** can be largely determined by the NE occurrences in the previous recent month (e.g., **T-1**). Hence it is reasonable to reduce the size of training corpus through dropping historical instances.

To predict the NEs in current month **T** based on data from month 1 to month $T-1$, I further break down the training corpus into weeks (from 1 to $N$), I introduce a coefficient $\delta$, as a means to defining a tunable parameter to cull redundant instances from the training data, reducing the dataset size, as follows:

$$\delta(E_k) = \frac{k}{N} * freq(E_k) \tag{5.1}$$

Here, the $k$ stands for the $k^{th}$ week, and the function $freq(E_k)$ returns the frequency of named entity $E$ in Week $k$. The resulting $\delta$ denotes the volume to be selected (randomly) as training corpus. Instances closer to the current time stamp is desired to keep more. This strategy is further validated through my experimentation, later in Section 5.2.

### 5.1.4 Problem Formalization

Having explained this, Figure 5.1 illustrates the example microblog post graphically, where the labels are given in squares. For the NER task, I again follow the widely-used **BIES** coding scheme where **B**, **I**, **E** and **S** stand for *beginning of a NE*, *inside a NE*, *end of a NE* and *single-word NE*, respectively. Tags of $T$ (*Title*), $P$ (*Production*) and $O$ ("outside"; non-NE word) are needed to distinguish the types of NE. The boundary words of NEs are labeled as $OL$ ("outside left"), $OR$("outside right") and $OS$("outside single") to assist the recognition of NE. Unlike the joint CWS/IWR problem that I earlier tackled in Chapter 3, this is a standard sequence labeling problem with only one (hidden) label to associate with each (observed) word.

### 5.1.5 LCRF Features

I use the same three broad feature classes – lexical, dictionary-based and statistical – to detect *Title* and *Product* labels, as was done for the joint CWS/IWR problem described earlier in Chapter 3. As for the modeling methodology, I choose word-based sequence labeling is employed due to its simplicity and robustness to label NEs with multiple words. Please note that while the feature classes are identical to those used for CWS/IWR, the individual feature types are specific to the problem of microtext NER.

**Lexical Features**. For a competitive baseline, I employ lexical (n-gram) features suggested by the previous work Liu et al. (2011b). These features are listed below [2]:

- Word 1-gram: $W_k (i-2 < k < i+2)$
- Word 2-gram: $W_k W_{k+1} (i-3 < k < i+2)$

[2] For notational convenience, I denote a word token as $W_i$. I use $W_{m:n}$ to express a subsequence starting at the position $m$ and ending at $n$. $len$ stands for the length of the subsequence, and $offset$ denotes the position offset from the current word $W_i$. I use $b$ (*beginning*), $m$ (*middle*) and $e$ (*ending*) to indicate the position of $W_k$ ($m \leq k \leq n$) within the string segment $W_{m:n}$.

- Character 1-gram: Prefix($W_k, i$),($0 < i < 2$)

  the $i-$characters prefix of $W_k$.

- Character 1-gram: suffix($W_k, i$),($0 < i < 2$)

  the $i-$characters prefix of $W_k$.

- Character 1-gram: Prefix($W_k - 1, i$),($0 < i < 2$)

  the $i-$characters prefix of $W_{k-1}$.

- Character 1-gram: suffix($W_k - 1, i$),($0 < i < 2$)

  the $i-$characters prefix of $W_{k-1}$.

- Word 1-gram Shape: $TP_k(i - 1 < k < i + 1)$

  A Chinese word can be assembled from Chinese characters, English letters, punctuations digits as well as the combination of them.

- Word 2-gram Shape: $TP_{k:k+1}(i - 1 < k < i + 1)$

  A bigram can be assembled from Chinese words, English letters, punctuations digits as well as the combination of them.

- Whether $W_k$ occurs in between paired punctuations

  This feature is used to capture the NEs that are delimited between paired indicator punctuations (*e.g.*, "《三国演义》" ["Three Kingdoms"]).

**Dictionary-based Features**. I propose new features (highlighted with "*") that indicate whether the input word sequence matches entries in certain lexica. I use the compiled boundary word list and NE list from the training instances as lexica, as described earlier. The final list of dictionary-based features employed are:

- (*) If $W_k$ $(i - 3 < k < i + 3)$ is a boundary word: *OL@k, OS@k or OR@k*

- (*) If $W_k$ $(i - 3 < k < i + 3)$ is a NE word: *b@k or e@k*

- If $W_{m:n}$ $(i-3<m<n<i+3, 0<n-m<4)$ matches one entry in NE list: *NE@m:n*; *len@offset*;

**Statistical Features**. I use a PMI variant (Church and Hanks, 1990) to model long-distance dependencies. Recall the example sentence from Figure 5.1. Aside from the boundary words, the *anchor* word "参加" ["attend"], is four words away from the *Title*. The anchor word is a potential hint that guides the recognition of "啦啦宝贝大PK" ["cheerleading PK show"]. In response to this, I determine *anchor* words through a ranking process utilizing PMI scores. The computed PMI values are further projected into five levels to feed the CRF model. For each character $W_k$, I incorporate the PMI with anchor words as follows:

- (*) Anchor Word 1-gram: *Anchor@offset*

  For example, in Figure 5.1, the feature for word "宝贝" ["cheerleader"] is 活动@3, in which "活动 ["show"]" is the anchor word.

- (*) Anchor Word 1-gram: *PMI@offset*

  The feature for word "宝贝" ["cheerleader"] is C@3, which is in the middle level.

## 5.2 Experiment

### 5.2.1 Baseline System

My baseline system uses only the **Lexical Feature** class. The resultant system uses similar surface lexical features used in previous state-of-the-art work Che et al. (2013). Noteworthy here is that the tags (*OL*, *OR* and *OS*) proposed to label boundary words of NEs are not employed in the baseline system. I use the open-source Mallet GRMM package to implement the sequential LCRF model.

### 5.2.2 Experimental Results

Figure 5.3 reports the performance of the baseline system trained on the training datasets from February to July 2011, and evaluated on the testing set of *TIT_July* and *PRO_July*. Each set contains 1000 annotated microblog posts. We can see that the performance dependence on recent training data is striking. Both figures show a strong upward trend as data up to the current month is incorporated.

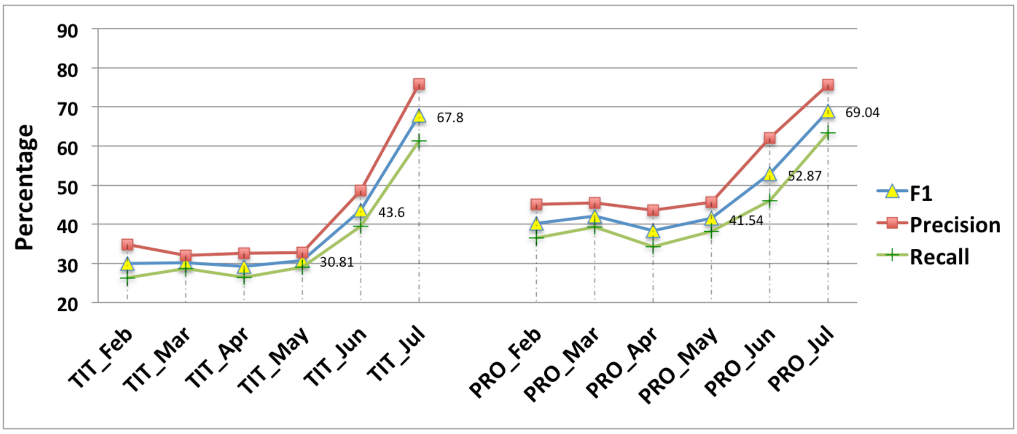


Figure 5.3: Baseline performance on the *Title* (left) and *Product* (right) recognition tasks. The $F_1$ score is labeled to the nearest data point.

According to Table 5.1, even incorporating several prior month's worth of training data (February to June) is worse than getting only one month's of recent training data (July only). It is clear that this validates my prior

Table 5.1: Baseline performance on the *Title* (upper) and *Product* (bellow) recognition tasks.

| Training Data Set | $F_1$ | PRE | REC |
|---|---|---|---|
| *TIT_Jun* | 43.60 | 48.63 | 39.51 |
| ***TIT_Jul*** | **67.80** | **75.78** | **61.34** |
| *TIT_Feb_to_Jun* | 52.81 | 60.11 | 47.09 |
| ***TIT_Feb_to_Jul*** | **71.19** | **76.95** | **66.23** |
| *PRO_Jun* | 52.87 | 61.97 | 46.10 |
| ***PRO_Jul*** | **69.04** | **75.75** | **63.42** |
| *PRO_Feb_to_Jun* | 56.54 | 62.85 | 51.23 |
| ***PRO_Feb_to_Jul*** | **72.65** | **78.15** | **67.87** |

observation that recent historical data is absolutely essential to performance. The improvement obtained by accumulating historical data for training is not as striking, but is statistically significant ($p < 0.05$) against the performance using data from the single most recent month.

I additionally observe that the performance on *Product* NER is better than *Title* recognition. This implies a longer lifecycle of *Product* NEs in the social media samples that my dataset represents from Weibo, compared to *Title*s.

Table 5.2 gives a comprehensive performance listing of several systems, varying feature sets (baseline or full), amount of training data (one or multiple months), use of balanced or raw data (+ negative), and NE task (*Product* or *Title*). From the results we see that there is consensus that la-

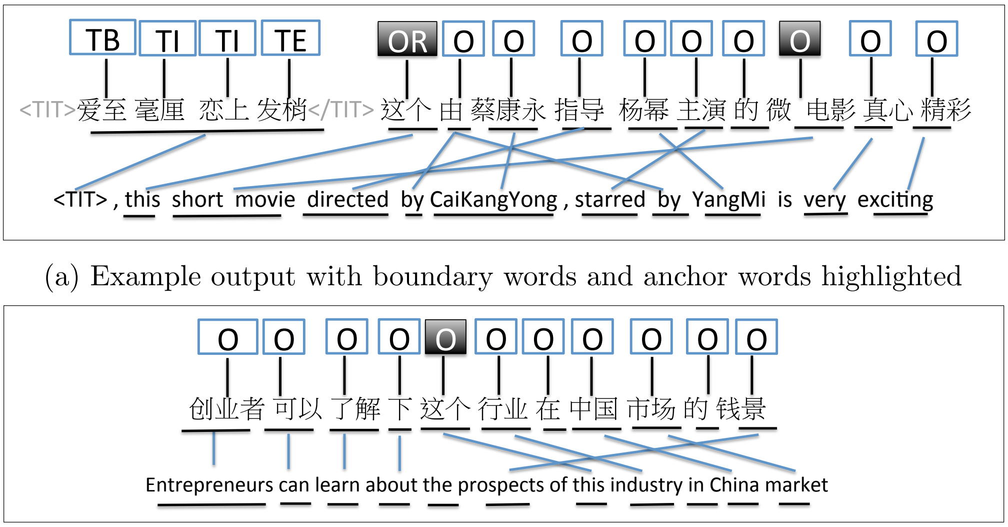
Table 5.2: $F_1$ comparison between systems with different feature sets and training corpus. The cells in bold highlight the difference compared to the previous row. **Negative** refers to the 2000 negative instances. **Full** denotes the system with boundary word labels and the three broad feature classes as a whole.

| | Training Data Set | System | $F_1$ |
|---|---|---|---|
| (1) | *TIT_Feb_to_Jun* | Baseline | 52.81 |
| (2) | *TIT_Feb_to_Jun* | **Full** | 55.81 |
| (3) | ***TIT_Feb_to_Jun + Negative*** | Full | **59.63** |
| (4) | *TIT_Jul* | Baseline | 67.8 |
| (5) | *TIT_Jul* | **Full** | 76.21 |
| (6) | ***TIT_Jul + Negative*** | Full | **78.24** |
| (7) | *TIT_Feb_to_Jul* | Baseline | 71.19 |
| (8) | *TIT_Feb_to_Jul* | **Full** | 78.34 |
| (9) | ***TIT_Feb_to_Jul + Negative*** | Full | **81.02** |
| (10) | *PRO_Feb_to_Jun* | Baseline | 56.45 |
| (11) | *PRO_Feb_to_Jun* | **Full** | 59.87 |
| (12) | ***PRO_Feb_to_Jun + Negative*** | Full | **62.49** |
| (13) | *PRO_Jul* | Baseline | 69.04 |
| (14) | *PRO_Jul* | **Full** | 73.00 |
| (15) | ***PRO_Jul + Negative*** | Full | **76.28** |
| (16) | *PRO_Feb_to_Jul* | Baseline | 72.65 |
| (17) | *PRO_Feb_to_Jul* | **Full** | 75.23 |
| (18) | ***PRO_Feb_to_Jul+ Negative*** | Full | **78.35** |

beling the boundary words by using evidence from the three broad feature classes improves $F_1$ scores. For example, one can see a marked increase of 9% in $F_1$ when comparing Row 4 to Row 5. This gives strong evidence of the importance of labeling boundary words, especially when the training set are up-to-date. Taking Figure 5.4a as an example, the baseline fails to recognize “爱至毫厘恋上发梢” as a *Title*, where the **Full** system succeeds. Inspecting the output, the enhanced system labels the word “这个” [“this”], as **OR**, with the help of anchor word “微电影” [“short movie”]. By considering the probabilities of the boundary words, the enhanced model infers better quality labels.

The effect of balancing the skewed training data can also be observed in the results. Balancing the training data to provide appropriate negative samples also improves performance. The rows with $F_1$ in bold in the figure validates utility of the negative training instances. For example, in cases where the label for the word “这个” [“this”], can exhibit an overly positive signal in the training corpus without negative instances, as shown in Figure 5.4a. By including the negative instance shown in Figure 5.4b, the bias caused by the positive signal is lessened.

To shed some insight into redundant instances filtering, Table 5.3 shows the effect of reducing the training size. Comparing the last row with the first row, a 45% reduction in training size is observed. The result highlighted in the last row suggests the similar performance is guaranteed with much a smaller training corpus leading to much a lighter model. This significantly saves the training and decoding time and potentially makes it possible to efficiently refresh the model with newly built corpus.


TB TI TI TE OR O O O O O O O O O
<TIT>爱至 毫厘 恋上 发梢</TIT> 这个 由 蔡康永 指导 杨幂 主演 的 微 电影 真心 精彩
<TIT> , this short movie directed by CaiKangYong , starred by YangMi is very exciting
O O O O O O O O O O O
创业者 可以 了解 下 这个 行业 在 中国 市场 的 钱景
Entrepreneurs can learn about the prospects of this industry in China market

(a) Example output with boundary words and anchor words highlighted

(b) A negative instance with the same word highlighted

Figure 5.4: Sample microtext with annotations for *Title*. The bottom line gives the aligned English translation.

Table 5.3: Performance comparison between systems with different training data sets.

| Training Data Set | $F_1$ | P | R | Model Size | Training Size |
|---|---|---|---|---|---|
| *TIT_Feb_to_Jun + Negative* | 59.63 | 65.86 | 54.47 | 338 (MB) | 5.8 (MB) |
| *TIT_Apr_to_Jun + Negative* | 58.02 | 63.98 | 53.07 | 212 (MB) | 3.6 (MB) |
| $\delta(TIT\_Feb\_to\_Jun)$ + ***Negative*** | **62.21** | **66.02** | **54.19** | **89** (MB) | **3.3** (MB) |

## 5.3 Summary

Starting from a building corpus automatically, I have demonstrated a time-sensitive approach for recognizing *Title* and *Product* NEs from microtext, leveraging anchor words and boundary words. Utilizing the same three broad feature classes – lexical, dictionary-based and statistical – my method addresses the observed shortcomings of standard supervised approaches of ambiguity and time-sensitivity. These features, alongside my approach that leverages 1) the intuition that the (long-distance) dependency between labeled boundary words and named entities matter, and 2) the distribution of named entity frequency against different time periods, show significant NER improvements over a competitive baseline. With these features, the improvements of 10% and 6% in $F_1$ for *Title* and *Product*, respectively, are obtained with the full training corpus.

A final contribution that I make is to propose an instance selection strategy to reduce the original model size by 75%, without hurting the recognition performance.

# Chapter 6

# Conclusion and Future Work

There has been increasing research attention paid to microtext, as it has proven a valuable data source for certain applications such as trend detection (Becker et al., 2011; Benhardus and Kalita, 2013; Mathioudakis and Koudas, 2010) and brand reputation monitoring (Gao et al., 2014; Guangyuan et al., 2011).

As the capability to process text in the formal domain has improved rapidly, researchers are increasingly focusing on other varieties of text, including the informal language found in microtext. proving the microtext as a valuable data source. The focus of my thesis is on bridging the language processing gap in such informal Chinese microtext. The ability to understand and process informal words and named entities from microtext has great potential to improve many downstream NLP applications, ranging from information filtering, sentiment analysis to machine translation.

## 6.1 Conclusion

In this thesis, I have motivated the importance of processing informal words and named entities from Chinese microtext. Accurate processing offers us an opportunity to glean more signal from the noisy corpora of microtext and offers a better understanding of the semantic underpinnings of microtext, and has great potential to benefit downstream applications.

Based on this motivation, my thesis has studied three key areas in building natural language processing tools for the microtext domain: 1) informal word recognition and word segmentation, 2) informal word normalization, and 3) named entity recognition.

To fill in the gap in existing literature, where less effort has been committed to explore the dependency between the informal words recognition and word segmentation, this work firstly proposes a joint inference model to tackel both problems simultaneously. It significantly out-performs the state-of-the-art performance on segmenting microtext by up to 8.0%. It is also able to recognize informal words from the microtext with the $F_1$ of 75.0%. To better understand the informal words, this thesis goes on to normalize informal words into formal counterparts. By incorporating rule-based and statistical features derived from both informal and formal text domain, together with proposed two-stage selection-classification algorithm, better normalization performance is observed. Besides the achievements verified by experiment results, I also build up the Chinese microtext corpus with crowdsourced annotations, which is publicly available for comparative research. To further study the informal elements – named entities – in microtext, this work presents an effective method to recognize the named entities frequently mentioned in microtext, using the corpus crawled and annotated automatically. This work identifies and addresses the difficulties of ambiguity, time-sensitivity as well as the concern of lacking corpus, and proposes the strategy to filter out redundant training instances.

With growing interest from NLP community, research on processing microtext – a representative type of UGC on social media – is getting more important and valuable.

## 6.2 List of Corpora and Tools

Here is the list of corpora and tools that have been built in this work. They are now publicly available and free to use for research purpose [1].

1. The Chinese microtext corpus with annotations of word segmentation and informal word recognition – a subset of *Dataset A* described in Chapter 3.2,

2. The Chinese microtext corpus with annotations of informal word normalization – a subset of *Dataset A* described in Chapter 4.1,

3. The Chinese microtext corpus with annotations of named entities – *Dataset B* described in Chapter 1,

4. A Chinese word segmenter and informal word detector [2] – presented in Chapter 3, and

5. A web-based lexicon presenting the informal words recognized automatically by the informal word detector [3].

## 6.3 Future Work

In this thesis, I have identified and conducted research in three important tasks in processing Chinese microtext. However, many remaining questions are worth of further exploration.

[1] The datasets are available at `http://wing.comp.nus.edu.sg/downloads/weiboCWSIWRData/`

[2] This tool is available at `http://wing.comp.nus.edu.sg:8080/CWSIWR/`

[3] This tool is available at `http://wing.comp.nus.edu.sg/weiboDemo/weibodict.html`

### 6.3.1 Word Segmentation on Chinese Microtext

The performance obtained for word segmentation can be further improved. As suggested in Section 3.3, By incorporating informal word normalization and named entity recognition, it is likely to improve the accuracy of recognition and segmentation. Besides the points noted from the error analysis, to speed up the development, I believe that it is important to gather a much bigger size of annotated corpus as well as to agree on a standard segmentation format. Such effort has been made in the recent SIGHAN shared tasks[4].

### 6.3.2 Informal Word Recognition and Normalization

As discussed in Section 4.1, the creation of informal words are motivated by different intents. In addition to those motivations with strong subjective views (*e.g.*, to be humorous or to avoid sensitive words), unconscious spelling error makes up another major group of informal words, which attracts the focus on Chinese spelling checking. The method proposed in this thesis is potentially enhanced-able to tackle the spelling error detection and correction in formal text domain.

Another possible direction is to cast normalization into a translation problem, which is to perform translation from informal text into formal text within the same language. Wang and Ng (2013) use 500 annotated sentence pairs to tune the SMT decoder, which takes the "hypothesis" (formal candidates) based on a manually assembled vocabulary containing 703 informal/formal pairs. The result evaluated by BLEU score is promising. I thus believe, it will be exciting to combine these efforts together with my work in candidates generation, so that the resulting system will not be

[4] `http://www.cipsc.org.cn/clp2012/task1.html`

limited by small but expensive vocabularies.

### 6.3.3 Named Entity Recognition on Chinese Microtext

In Chapter 5, I have noted that the use of boundary words and anchor words can help enhance the sentence-wide information that can be captured. One key enhancement is to model the long-distance dependencies among them. Since long-distance dependencies are difficult to represent in generative models, the skip-chain CRF model here could potentially be employed as it takes the advantage of flexibility in allowing input-specific model structure.

It is also important to explore other ways to obtain negative training instances. As we depart from leveraging punctuation and dictionary entries to label entities, the drawback of resulting a biased corpus becomes well-marked. Another point is that training instances filtering at the moment plays a moderate role in improving the performance of recognition. There is a big room to further study how the named entities are trending along with the time and also to better formulate the deduction strategy.

## 6.4 List of published works

Here is the list of published works during my Ph.D. candidature. Works closely related to this thesis are highlighted by "(*)". The recent research works focusing on Chinese microtext benefit greatly from my previous research effort committed on studying English tweets and annotation crowdsourcing.

1. (*) Aobo Wang, Min-Yen Kan, Daniel Andrade, Takashi Onishi, and Kai Ishikawa. Chinese informal word normalization: an experimental

study. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing, IJCNLP*, volume 13, pages 127–135, 2013.

2. (*) Aobo Wang and Min-Yen Kan. Mining informal language from chinese microtext: Joint word recognition and segmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–741, 2013.

3. Aobo Wang, Tao Chen, and Min-Yen Kan. Re-tweeting From A Linguistic Perspective. In *Proceedings of the Second Workshop on Language in Social Media*, pages 46–55, 2012a.

4. Aobo. Wang, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives on Crowdsourcing Annotations for Natural Language Processing, journal = Language Resources and Evaluation. pages 1–23, 2010.

# Bibliography

P. Achananuparp, I.N. Lubis, Y. Tian, D. Lo, and E.P. Lim. Observatory of Trends in Software Related Microblogs. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, pages 334–337. ACM, 2012.

Masayuki Asahara, Chooi Ling Goh, Xiaojie Wang, and Yuji Matsumoto. Combining segmenter and chunker for chinese word segmentation. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 144–147. Association for Computational Linguistics, 2003.

AiTi Aw, Min Zhang, Juan Xiao, and Jian 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. URL `http://www.aclweb.org/anthology/P/P06/P06-2005`.

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédrick 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*, pages 770–779. Association for Computational Linguistics, 2010. URL `http://www.aclweb.org/anthology/P10-1079`.

Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11:438–441, 2011.

Yassine Benajiba, Imed Zitouni, Mona Diab, and Paolo Rosso. Arabic named entity recognition: Using features extracted from noisy data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 281–285, 2010.

James Benhardus and Jugal Kalita. Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1):122–139, 2013.

Aditya Bhargava and Grzegorz Kondrak. How do you pronounce your name?: improving g2p with transliterations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 399–408, 2011.

Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

Thorsten Brants and Alex Franz. The google web 1t 5-gram corpus version 1.1. *LDC2006T13*, 2006.

Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570. Association for Computational Linguistics, 2011. URL `http://www.aclweb.org/anthology/D11-1052`.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S

Roossin. A statistical approach to machine translation. *Computational linguistics*, pages 79–85, 1990.

Belinda Cao. Sina's weibo outlook buoys internet stock gains: China overnight, 2012. URL `http://www.bloomberg.com/news/2012-02-28/sina-s-weibo-outlook-buoys-internet-stock-gains-in-n-y-china-overnight.html`.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, pages 27:1–27:27, 2011.

Jing-Shin Chang and Wei-Lun Teng. Mining atomic chinese abbreviations with a probabilistic single character recovery model. *Language Resources and Evaluation*, pages 367–374, 2006.

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. Named entity recognition with bilingual constraints. In *Proceedings of NAACL-HLT*, pages 52–62, 2013.

Keh-Jiann Chen and Ming-Hong Bai. Unknown word detection for chinese by a corpus-based learning method. *International Journal of Computational Linguistics and Chinese Language Processing*, 3(1):27–44, 1998.

Keh-Jiann Chen and Wei-Yun Ma. Unknown Word Extraction for Chinese Documents. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, 2002.

T. Chen and M.Y. Kan. Creating a Live, Public Short Message Service Corpus: the NUS SMS Corpus. *Language Resources and Evaluation*, pages 1–37, 2011.

Chen Cheng, Xian Yi Cheng, and Jin Hua. Research of chinese named entity recognition using gate. *Advanced Materials Research*, 393:262–264, 2012.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012, 2010.

Kenneth Ward Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computional Linguistic*, pages 22–29, 1990.

Anqi Cui, Liner Yang, Dejun Hou, Min-Yen Kan, Yiqun Liu, Min Zhang, and Shaoping Ma. PrEV: Preservation Explorer and Vault for Web 2.0 User-Generated Content. *Theory and Practice of Digital Libraries*, pages 101–112, 2012.

E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Towards Personalized Learning to Rank for Epidemic Intelligence Based on Social Media Streams. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 495–496. ACM, 2012.

Marco Dinarelli and Sophie Rosset. Models cascade for tree-structured named entity detection. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1269–1278, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/I11-1142`.

Junwu Du, Zhimin Zhang, Jun Yan, Yan Cui, and Zheng Chen. Using search session context for named entity recognition in query. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 765–766, 2010.

Huanzhong Duan and Yan Zheng. A study on features of the crfs-based chinese named entity recognition. *International Journal of Advanced Intelligence*, 3(2):287–294, 2011.

Jenny Rose Finkel and Christopher D Manning. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 141–150. Association for Computational Linguistics, 2009.

Jenny Rose Finkel and Christopher D. Manning. Hierarchical joint learning: improving joint parsing and named entity recognition with non-jointly labeled data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 720–728, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1858681.1858755`.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, 2005.

S. Fujiki, H. Yano, T. Fukuda, and H. Yamana. Retweet Reputation: A Bias-Free Evaluation Method for Tweeted Contents. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. Chinese Word

Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computitional Linguistic*, pages 531–574, 2005.

Yue Gao, Fanglin Wang, Huanbo Luan, and Tat-Seng Chua. Brand data gathering from live social media streams. In *Proceedings of International Conference on Multimedia Retrieval*, page 169. ACM, 2014.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. Contextual Bearing on Linguistic Variation in Social Media. In *Proceedings of the Workshop on Language in Social Media*, pages 20–29, 2011.

Li Guangyuan, Cao Jinping, Jiang Jing, Li Qian, and Yao Ling. Brand tweets: how to popularize the enterprise micro-blogs. In *Information Technology and Artificial Intelligence Conference (ITAIC), 2011 6th IEEE Joint International*, volume 1, pages 136–139. IEEE, 2011.

Zhao Hai, Huang Chang-Ning, Li Mu, and Lu Bao-Liang. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. *The 20th Pacific Asia Conference on Language, Information and Computation*, pages pp.87–94, 2006.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *ACM Special Interest Groups on Knowledge Discovery and Data Mining Explorations Newsletter*, pages 10–18, 2009.

Bo Han and Timothy 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, 2011.

Bo Han, Paul Cook, and Timothy 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*, pages 421–432, 2012.

Chang-Ning Huang and Hai Zhao. Chinese word segmentation: A decade review. *Journal of Chinese Information Processing*, 21(3):8–20, 2007.

Finn V Jensen. *An Introduction to Bayesian Networks*, volume 74. 1996.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, volume 2007, page 22, 2007.

Fleiss L Joseph. Measuring nominal scale agreement among many raters. *Psychological bulletin*, pages 378–382, 1971.

Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pages 1–8, 2002.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 694–702, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P12-1073`.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, pages 180–183, 2003.

Catherine Kobus, François Yvon, and Géraldine Damnati. Normalizing SMS: Are Two Metaphors Better Than One? In *International Conference on Computational Linguistics*, pages 441–448, 2008.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730. ACM, 2012.

Hongqiao Li, Chang-Ning Huang, Jianfeng Gao, and Xiaozhong Fan. The use of svm for chinese new word identification. In *Natural Language Processing–IJCNLP 2004*, pages 723–732. Springer, 2005.

Zhifei Li and David 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*, pages 1031–1040, 2008a.

Zhifei Li and David Yarowsky. Unsupervised translation induction for chinese abbreviations using monolingual corpora. In *Proceedings of ACL*, pages 425–433, 2008b.

Zhongguo Li and Maosong Sun. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512, 2009.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. Insertion, Deletion, or Substitution? Normalizing Text Messages without Pre-categorization nor Supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76. Association for Computational Linguistics, 2011a. URL `http://www.aclweb.org/anthology/P11-2013`.

Fei Liu, Fuliang Weng, and Xiao Jiang. A broadcoverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1035–1044, 2012.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, 2011b.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158. ACM, 2010.

James Mayfield, Paul McNamee, and Christine Piatko. Named entity recognition using hundreds of thousands of features. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 184–187, 2003.

Hye-Jin Min and Jong C Park. Product name classification for product instance distinction. *Pacific Asia Conference on Language, Information, and Computation*, 2012.

Robert Munro and Christopher D. Manning. Accurate unsupervised joint named-entity extraction from unaligned parallel text. In *Proceedings of the 4th Named Entity Workshop*, NEWS '12, pages 21–29, 2012.

K. Nishida, T. Hoshide, and K. Fujimura. Improving tweet stream classification by detecting changes in word probability. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 971–980. ACM, 2012.

Sebastian Padó. *User's Guide to* **`SIGF`***: Significance Testing by Approximate Randomisation*, 2006.

Serguei Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 160–167, 2002.

Youngja Park and Roy J Byrd. Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the 2001 conference on empirical methods in natural language processing*, pages 126–133, 2001.

Adam Pauls and Dan Klein. Faster and smaller n-gram language models. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 258–267, 2011.

Fuchun Peng and Dale Schuurmans. Self-supervised chinese word segmentation. In *Advances in Intelligent Data Analysis*, pages 238–247. Springer, 2001.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. Chinese Segmentation and New Word Detection Using Conditional Random Fields. In *Proceedings of the 20th international conference on Computational Linguistics*, page 562, 2004.

Xipeng Qiu, Qi Zhang, and Xuanjing Huang. Fudannlp: A toolkit for chinese natural language processing. In *ACL (Conference System Demonstrations)*, pages 49–54. Citeseer, 2013.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, 2009.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1141`.

Gökhan Akin Seker and Gülsen Eryigit. Initial explorations on using crfs for turkish named entity recognition. In *COLING*, pages 2459–2474, 2012.

Satoshi Sekine. Named entity: History and future. *Proteus project report*, 2004.

S. Song, Q. Li, and X. Zheng. Detecting popular topics in micro-blogging based on a user interest-based model. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*, pages 841–842. ACM, 2010.

Weiwei Sun and Jia Xu. Enhancing Chinese Word Segmentation Using Unlabeled Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 970–979, 2011.

Xu Sun, Houfeng Wang, and Wenjie Li. Fast Online Training with Frequency-Adaptive Learning Rates for Chinese Word Segmentation and New Word Detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 253–262, 2012.

Charles Sutton. GRMM: GRaphical Models in Mallet. In *URL http://mallet. cs. umass. edu/grmm*, 2006.

Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. *Journal of Machine Learning Research*, pages 693–723, 2007.

Stephanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 41–47, 2003.

Xiaojun Wan, Liang Zong, Xiaojiang Huang, Tengfei Ma, Houping Jia, Yuqian Wu, and Jianguo Xiao. Named entity recognition in chinese news comments on the web. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 856–864, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/I11-1096`.

Aobo Wang and Min-Yen Kan. Mining informal language from chinese microtext: Joint word recognition and segmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–741, 2013.

Aobo. Wang, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives on Crowdsourcing Annotations for Natural Language Processing, journal = Language Resources and Evaluation. pages 1–23, 2010.

Aobo Wang, Tao Chen, and Min-Yen Kan. Re-tweeting From A Linguistic Perspective. In *Proceedings of the Second Workshop on Language in Social Media*, pages 46–55, 2012a.

Aobo Wang, Min-Yen Kan, Daniel Andrade, Takashi Onishi, and Kai Ishikawa. Chinese informal word normalization: an experimental study. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing, IJCNLP*, volume 13, pages 127–135, 2013.

Longyue Wang, Shuo Li, Derek F Wong, and Lidia S Chao. A joint chinese named entity recognition and disambiguation sys-tem. In *The 2nd CIPSSIGHAN Joint Conference on Chinese Language Processing (CLP-2012)*, 2012b.

Pidong Wang and Hwee Tou Ng. A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of NAACL-HLT*, pages 471–481, 2013.

Xuancong Wang, Hwee Tou Ng, and Khe Chai Sim. Dynamic conditional random fields for joint sentence boundary and punctuation prediction. In *International Speech Comunication Association*, 2012c.

Kam-Fai Wong and Yunqing Xia. Normalization of Chinese Chat Language. *Language Resources and Evaluation*, pages 219–242, 2008.

Andi Wu and Zixin Jiang. Statistically-Enhanced New Word Identification in A Rule-based Chinese Aystem. In *Proceedings of the second workshop on Chinese Language Processing*, pages 46–51, 2000.

Dan Wu, Wee Sun Lee, Nan Ye, and Hai Leong Chieu. Domain adaptive bootstrapping for named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1523–1532. Association for Computational Linguistics, 2009.

Jian-Cheng Wu and Jason S Chang. Learning to find english to chinese transliterations on the web. In *Proc. of EMNLP-CoNLL*, pages 996–1004, 2007.

Yunqing Xia and Kam-Fai Wong. Anomaly Detecting within Dynamic Chinese Chat Text. *NEW TEXT Wikis and blogs and other dynamic text sources*, page 48, 2006.

Yunqing Xia, Kam-Fai Wong, and Wei Gao. NIL Is Not Nothing: Recognition of Chinese Network Informal Language Expressions. In *4th SIGHAN Workshop on Chinese Language Processing*, volume 5, 2005.

Yunqing Xia, Kam-Fai Wong, and Wenjie Li. A phonetic-based approach to chinese chat text normalization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 993–1000, 2006.

Nianwen Xue. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*, pages 29–48, 2003.

Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, pages 947–953, 2000.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 184–187, 2003.

Min Zhang, Xiangyu Duan, Vladimir Pervouchine, and Haizhou Li. Machine transliteration: leveraging on third languages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1444–1452, 2010.

Wei Zhang, Yan Chuan Sim, Jian Su, and Chew Lim Tan. Entity linking with effective acronym expansion, instance selection, and topic modeling. In *International Joint Conference on Artificial Intelligence*, volume 2011, pages 1909–1914, 2011.

Hai Zhao and Chunyu Kit. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*, pages 106–111, 2008.

GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480, 2002.

Arkaitz Zubiaga, Damiano Spina, Víctor Fresno, and Raquel Martínez. Classifying trending topics: a typology of conversation triggers on Twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2461–2464. ACM, 2011. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063992. URL `http://doi.acm.org/10.1145/2063576.2063992`.