Mining Informal Language from Chinese Microtext: Joint Word Recognition and Segmentation

Aobo Wang[1] **Min-Yen Kan**[1,2*]

[1] Web IR / NLP Group (WING)
[2] Interactive and Digital Media Institute (IDMI)
National University of Singapore
13 Computing Link, Singapore 117590
{wangaobo,kanmy}@comp.nus.edu.sg

Abstract

We address the problem of informal word recognition in Chinese microblogs. A key problem is the lack of word delimiters in Chinese. We exploit this reliance as an opportunity: recognizing the relation between informal word recognition and Chinese word segmentation, we propose to model the two tasks jointly. Our joint inference method significantly outperforms baseline systems that conduct the tasks individually or sequentially.

1 Introduction

User generated content (UGC) – including microblogs, comments, SMS, chat and instant messaging – collectively referred to as *microtext* by Gouwset *et al.* (2011) or *network informal language* by Xia *et al.* (2005), is the hallmark of the participatory Web.

While a rich source that many applications are interested in mining for knowledge, microtext processing is difficult to process. One key reason for this difficulty is the ubiquitous presence of informal words – anomalous terms that manifest as *ad hoc* abbreviations, neologisms, unconventional spellings and phonetic substitutions. Such informality is often present in oral conversation, and user-generated microblogs reflect this informality. Natural language processing (NLP) tools largely fail to work properly on microtext, as they have largely been trained on formally written text (i.e., newswire). Recent work has started to address these shortcomings (Xia and Wong, 2006; Kobus et al., 2008; Han and Baldwin, 2011). Informal words and their usage in microtext evolves quickly, following social trends and news events. These characteristics make it difficult for lexicographers to compile lexica to keep with the pace of language change.

We focus on this problem in the Chinese language. Through our analysis of a gathered Chinese microblog corpus, we observe that Chinese informal words originate from three primary sources, as given in Table 1.

But unlike noisy words in English, Chinese informal words are more difficult to mechanically recognize for two critical reasons: first, Chinese does not employ word delimiters; second, Chinese informal words combine numbers, alphabetic letters and Chinese characters. Techniques for English informal word detection that rely on word boundaries and informal word orthography need to be adapted for Chinese. Consider the microtext "不要剧透了" (meaning "Don't tell me the spoilers (to a movie or joke)", also in Table 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[1] incorrectly yield "不要␣剧␣透了", preferring to chunk "透了" ("thoroughly") as a word, as they do not consider the possibility that "剧透" ("spoiler") could be an informal word. This example illustrates the mutual dependency between Chinese word segmentation (henceforth, CWS) and informal word recognition (IWR) that should be solved jointly.

Hence, rather than pipeline the two processes serially as previous work, we formulate it as a two-layer sequential labeling problem. We employ factorial conditional random field (FCRF) to solve both CWS and IWR jointly. To our best knowledge, this is the first work that shows how Chinese microtext can be analyzed from raw text to

*This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

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

Table 1: Our classification of Chinese informal words as originating from three primary sources. For **Phonetic Substitutions**, pronunciation is indicated by the phonetic Pinyin transcription system.

	Informal Word	**Formal Word**	**Example Sentence**	**English Translation**
1) Phonetic Substitutions	木有(**mu4** you3)	没有(**mei2** you3)	开发区**木有**出租车	**No** taxi in the development area
	孩纸们(hai2 **zhi3** men)	孩子们(hai2 **zi** men)	起床了**孩纸们**	Get up **kids**
	bs	鄙视(**bi shi**)	我**bs**你	I **despise** you
2) Abbreviation	桌游	桌面游戏	来**桌游**吧	Let’s play board games
	剧透	剧情透露	不要**剧透**了	Don’t **tell (me) the spoilers**
3) Neologisms	给力	很棒	真**给力**啊	So **awesome**!
	秒杀	迅速购买	速度**秒杀**它	**Quickly purchase** it

derive joint solutions for both problems of CWS and IWR. We also propose novel features for input to the joint inference. Our techniques significantly outperform both research and commercial state-of-the-art for these problems, including two-step linear CRF baselines which perform the two tasks sequentially.

We detail our methods in Section 2. In Section 3, we first describe the details of our dataset and baseline systems, followed by demonstrating two sets of experiments for CWS and IWR, respectively. Section 4 offers the discussion on error analysis and limitations. We discuss related work in Section 5, before concluding our paper.

2 Methodology

Given an input Chinese microblog post, our method simultaneously segments the sentences into words (the Chinese Word Segmentation, CWS, task), and marks the component words as informal or formal ones (the Informal Word Recongition, IWR, task).

2.1 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, we follow the widely-used **BIES** coding scheme (Low et al., 2005; Hai et al., 2006), 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, we have two (hidden) labels to associate with each (observable) character. Figure 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”).

2.2 Conditional Random Field Models

Given the 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, we feel that for this task, formulating efficient global formulas (constraints) for MLN (ILP) is comparatively 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.

2.2.1 Linear-Chain CRF

A linear-chain CRF (LCRF; Figure 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.

2.2.2 Factorial CRF

To properly model the interplay between the two sub-problems, we 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 captures the joint distribution among various layers and jointly predicts across layers. Figure 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.

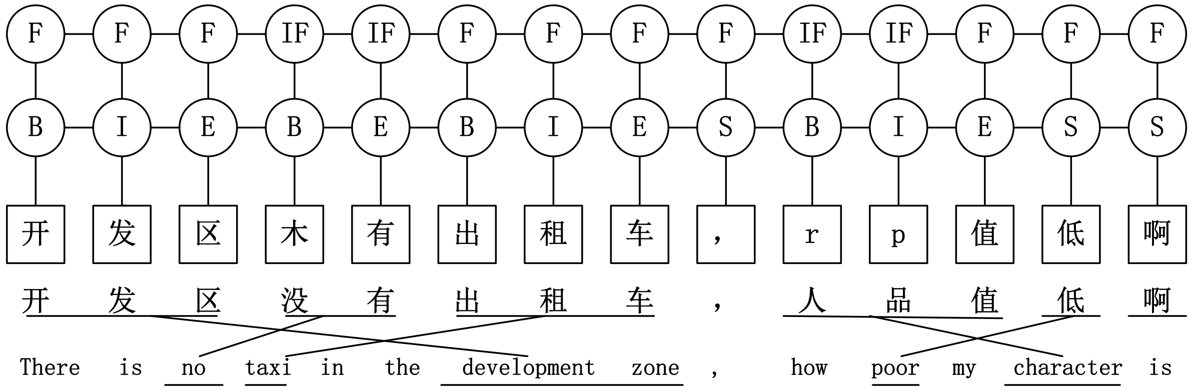
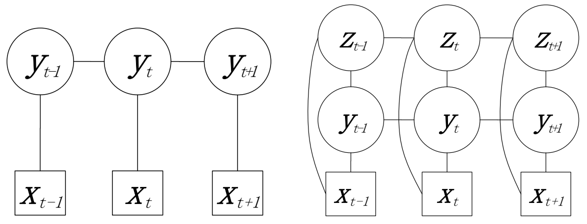

F F F IF IF F F F F IF IF F F F
B I E B E B I E S B I E S S
开 发 区 木 有 出 租 车 ， r p 值 低 啊
开 发 区 没 有 出 租 车 ， 人 品 值 低 啊
There is no taxi in the development zone , how poor my character is

Figure 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 pronuniciation in Pinyin and aligned English translation.



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

Figure 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.

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 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, we follow the FCRF model, using exact Junction Tree (Jensen, 1996) inference and decoding algorithm to perform prediction.

2.3 CRF Features

We 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 our work is also to propose novel features for joint inference. We propose new features for the dictionary-based and statistical feature classes, which we have marked in the discussion below with "(*)". We later examine their efficacy in Section 3.

Lexical Features. As a foundation, we employ lexical (n-gram) features informed by the previous state-of-the-art for CWS (Sun and Xu, 2011; Low et al., 2005). These features are listed below[2]:

- 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. We use features that indicate whether the input character sequence

[2] For notational convenience, we denote a candidate character token C_i as having a context $...C_{i-1}C_iC_{i+1}...$. We 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. We use b (*beginning*), m (*middle*) and e (*ending*) to indicate the position of C_k ($m \leq k \leq n$) within the string $C_{m:n}$.

matches entries in certain lexica. We use the online dictionary from Peking University as the formal lexicon and the compiled informal word list from our training instances as the informal lexicon. In addition, we employ additional online word lists[3] to distinguish named entities and function words from potential informal words.

As shown in Table 1, alphabetic sequences in microblogs may refer to Chinese Pinyin or Pinyin abbreviations, rather than English (e.g., "*bs*" for *bi shi*; "to despise"). Hence, we added dictionary-based features to indicate the presence of Pinyin initials, finals and standard Pinyin expansions, using a UK English word list[4]. 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. We 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 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 these difficulties in differentiating linguistic registers, we compute two different PMI scores for character-based bigrams from two large corpora representing news and microblogs as features. We also use the difference between the two PMI scores as a differential feature. In addition, we also convert all the character-based bigrams into Pinyin-based bigrams (ignoring tones[5]) 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 we eschew smoothing in our 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, we 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*

3 Experiment

We discuss the dataset, baseline systems and experiments results in detail in the following.

3.1 Data Preparation

We utilize the Chinese social media archive, PrEV (Cui et al., 2012), to obtain Chinese mi-

[3] Resources are available at `http://www.sogou.com/labs/resources.html`

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

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

croblog posts from the public timeline of Sina Weibo[6]. 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. Our dataset has a total of 6,678,021 messages, covering two months from June to July of 2011. To annotate the corpus, we employ *Zhubajie*[7], one of China mainland's largest crowdsourcing (Wang et al., 2010) platforms to obtain informal word annotations. In total, we 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' κ (Joseph, 1971), and conflicts were resolved by majority voting.

We divided the annotated corpus, taking $4,000$ posts for training, and the remainder ($1,500$) for testing. Through inspection, we note that 79.8% of the informal words annotated in the testing set are not covered by the training set. We also follow Wang *et al.* (2012)'s conventions and apply rulesets to preprocess the corpus' *URLs*, *emoticons*, *"@usernames"* and *Hashtags* as pre-segmented words, before input to CWS and IWR. For the CWS task, the first author manually labelled the same corpus following the segmentation guidelines published with the *SIGHAN-5*[8] MSR dataset.

3.2 Baseline Systems

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

Existing Systems. We re-implemented Xia and Wong (2008)'s extended Support Vector Machine (SVM) based microtext IWR system to compare with our method. Their system only does IWR, using the CWS and POS tagging otuput of the ICTCLAS segmenter (Zhang et al., 2003) as input. To compare our joint inference versus other learning models, we also employed a decision tree (DT) learner, equipped with the same feature set as our 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, we compare with two recent segmenters. Sun and Xu (2011)'s work achieves 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, we 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, we compare with a LCRF solution that chains these two tasks together. For completeness, we test pipelining in both directions – CWS feeding features for IWR ($LCRF_{cws} \succ LCRF_{iwr}$), and the reverse ($LCRF_{iwr} \succ LCRF_{cws}$). We modify the open-source Mallet GRMM package (Sutton, 2006) to implement both this sequential LCRF model and our proposed FCRF model. Both models take the whole feature set described in Section 2.3.

Upper Bound Systems. To measure the upper-bound achievable with perfect support from the complementary task, we 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$\succ$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, we also compared our 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). We 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).

3.3 Evaluation Metrics

We 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, we employ the conventional scoring script provided in *SIGHAN-*

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

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

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

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

5, which also provides out-of-vocabulary recall (OOVR).

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

3.4 Experimental Results

The goal of our 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 we 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 our proposed FCRF?

3.4.1 CWS Performance

Table 2: 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**	**$\mathbf{F_1}$**	**OOVR**
ICTCLAS (2003)	0.640	0.767	0.698	0.551
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^*	$0.758^‡$

In general, our FCRF yields the best performance among all systems (top portion of Table 2), answering RQ1. Given microblog posts as test data, the 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 shared by the two tasks, CWS benefits significantly from the results of IWR. Hence our segmenter outperforms the existing segmenters by tackling one of the bottlenecks of recognizing informal words in Chinese microtext.

To illustrate, the sequence "...有木有人..." ("...有没有人..."; "...is there anyone..."), is correctly labeled as **BIES** by our 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 we conclude that jointly modeling both tasks is more effective than performing the tasks sequentially.

For RQ3, the last two rows presents 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 our assumption that CWS can benefit from joint consideration of IWR. Taking the best previous work as our lower bound (0.69 F_1), we see that our FCRF methodology (0.77) makes significant progress towards the upper bound (0.82).

3.4.2 IWR Performance

For RQ1 and RQ2, Table 3 compares the performance of our method with the baseline systems on the IWR task. Overall, the FCRF method again outperforms all the baseline systems. We note 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 phe-

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

Table 3: 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}$$\succ$**LCRF**$_{iwr}$	0.858‡	0.591‡	0.699*
FCRF	**0.877***	**0.655***	**0.750***
LCRF$_{cws}$$\succ$**LCRF**$_{iwr}$**-UB**	0.840	0.726*	0.779*
FCRF-UB	0.878	0.752*	0.810*

nomenon more closely, we 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".

As indicated by the bold figures in Table 3, 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 we expect on real-world data that CWS performance will not be perfect. As such, we still recommend using the FCRF model, as the joint process is more robust to noisy input from one channel.

Table 4: 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*

3.4.3 Feature set evaluation

For RQ4, to evaluate the effectiveness of our newly-introduced feature sets (those marked with "*" in Section 2.3), we also test a FCRF (**FCRF**$_{-new}$) without our new features. According to Table 4, performance drops by a significant amount: 0.088 F$_1$ on CWS and 0.198 F$_1$ on IWR. **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.

3.4.4 Adapted SVM-JC vs. FCRF

Table 5: 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*

For RQ5, according to Table 5, our SVM trained to predict the cross-product CWS/IWR classification (SVM-JC) performs quite well on its own. Unsurprisingly, it does not outperform our 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 we think is partially explained by its good performance (0.761) on the CWS task. The over-prediction tendency of the individual SVM is largely solved 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 ($p < 0.05$).

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

4 Discussion

We wish to understand the causes of errors in our models so that we may better understand its weaknesses. Manually inspecting the errors of our system, we found three major categories of errors which we dissect here.

For IWR, the major source of error, accounting for more than 60% of all errors, is caused by what we 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 we 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). We can 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.

Table 6: 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

Another major group of errors come from what we term *freestyle named entities* as exemplified in Table 6; 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 Name* (PER), as defined in CoNLL-2003[11] Named Entity Recognition shared task. Such freestyle entities are often misrecognized 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. We recognize this as a challenge in Chinese microtext, but beyond the scope of our current work.

5 Related Work

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

In processing Chinese informal language, work conducted by Xia and Wong address the problem

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

of in bulletin board system (BBS) chats. They employ pattern matching and SVM-based classification to recognize Chinese informal sentences (not individual words) chat (Xia et al., 2005). Both methods had 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, our work identifies individual informal words, a finer-grained (and more difficult) task.

While seminal, we 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 our study covering almost triple the word types $(1,658)$. Our corpus 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 ours.

More closely related, Li and Yarowsky (2008) tackle Chinese IWR. They bootstrap 500 informal/formal word pairs by using manually-tuned queries to find definition sentences on the Web. The resulting noisy list is further re-ranked based on n-gram co-occurrence. However, their method makes a basic assumption that informal/formal word pairs co-occur within a definition sentence (i.e., "<*informal word*> means <*formal word*>") may not hold in microblog data, as microbloggers largely do not define the words they use.

Closely related to our work is the task of Chinese new word detection, normally treated as a separate process from word segmentation in most previous works (Chen and Bai, 1998; Wu and Jiang, 2000; Chen and Ma, 2002; Gao et al., 2005). Aiming to improve both tasks, work by Peng *et al.* (2004) and Sun *et al.* (2012) conduct 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. We note that the task of new word detection refers to out-of-vocabulary (OOV) detection, and is distinctly different from IWR (new words could be both formal or informal words).

6 Conclusion

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

We propose novel features including statistical and lexical features that improve the performance of the inference process. We evaluate our method on a manually-constructed data set and compare it with multiple research and industrial baselines that perform CWS and IWR individually or sequentially. Our experimental results show our joint inference model yields significantly better F_1 for both tasks. For analysis, we 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 our system, we 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. We believe incorporating informal word normalization into the inference process may help address this important source of error.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. We also appreciate the proofreading effort made by Tao Chen, Xiangnan He, Ning Fang, Yushi Wang and Haochen Zhan from WING. This work also benefits from the discussion with Yang Liu, associate professor from Tsinghua University.

References

Belinda Cao. 2012. Sina's weibo outlook buoys internet stock gains: China overnight.

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

Keh-Jiann Chen and Ming-Hong Bai. 1998. Unknown Word Detection for Chinese by a Corpus-Based Learning Method. *International Journal of Computational Linguistics and Chinese Language Processing*, pages 27–44.

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

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

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

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computitional Linguistic*, pages 531–574.

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

Zhao Hai, Huang Chang-Ning, Li Mu, and Lu Bao-Liang. 2006. 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.

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

Bo Han and Timothy Baldwin. 2011. 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.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. 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.

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

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

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

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

Zhifei Li and David Yarowsky. 2008. 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.

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

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

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

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

Xu Sun, Houfeng Wang, and Wenjie Li. 2012. 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.

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

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

Stephanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. 2003. 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.

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

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

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

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

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

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

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

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

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

Hai Zhao and Chunyu Kit. 2008. 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.