Chinese Informal Word Normalization: an Experimental Study

Aobo Wang[1]*, **Min-Yen Kan**[1,2]†, **Daniel Andrade**[3], **Takashi Onishi**[3], **Kai Ishikawa**[3]
[1] Web IR / NLP Group (WING)
[2] Interactive and Digital Media Institute (IDMI)
National University of Singapore
{wangaobo,kanmy}@comp.nus.edu.sg
[3] Knowledge Discovery Research Laboratories
NEC Corporation, Nara, Japan
{s-andrade@cj,t-onishi@bq,k-ishikawa@dq}.jp.nec.com

Abstract

We study the linguistic phenomenon of informal words in the domain of Chinese microtext and present a novel method for normalizing Chinese informal words to their formal equivalents. We 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 and informal pairs. Our two-stage selection-classification model is evaluated on a crowdsourced corpus and achieves a normalization precision of 89.5% across the different channels, significantly improving the state-of-the-art.

1 Introduction

Microtext – 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 receiving a larger research focus from the computational linguistic community. A key challenge is the presence of *informal words* – terms that manifest as *ad hoc* abbreviations, neologisms, unconventional spellings and phonetic substitutions. This phenomenon is so prevalent a challenge in Chinese microtext that the dual problems of informal word recognition and normalization deserve research. Given the close connection between an informal word and its formal equivalent, the restoration (normalization) of an informal word to its formal one is an important pre-processing step for NLP tasks that rely on string matching or word frequency statistics (Han et al., 2012).

It is important 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 informal 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"). It is thus desirable to normalize informal words to their standard formal equivalents before proceeding with standard text processing workflows.

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

We evaluate our system on a crowdsourced corpus, achieving good performance with a normalization precision of 89.5%. We also show that the method can be effectively adapted to tackle the synonym acquisition task in the formal domain. To our best knowledge, this is the first work to systematically explore the informal word phenomenon in Chinese microtext. By using a formal domain corpus, we introduce a method that effectively normalizes Chinese informal words through

*This research is done in part during Aobo Wang's internship in NEC Corporation.

†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.

different, independent channels.

2 Related Work

Previous works that address a similar task includes the study on abbreviations with their definitions (e.g., (Park and Byrd, 2001; Chang and Teng, 2006; Li and Yarowsky, 2008b)), abbreviations and acronyms in medical domain (Pakhomov, 2002), and transliteration (e.g., (Wu and Chang, 2007; Zhang et al., 2010; Bhargava and Kondrak, 2011)). These works dealt with such relations in formal text, but as we earlier argued, similar processing in the informal domain is quite different.

Probably the most related work to our method is Li and Yarowsky (2008a)'s work. They tackle the problem of identifying informal–formal Chinese word pairs in the Web domain. They employ the Baidu[1] search engine to obtain definition sentences – sentences that define or explain Chinese informal words with formal ones – from which the pairs are extracted and further ranked using a conditional log-linear model. Their method only works for definition sentences, where the assumption that the formal and informal equivalents co-occur nearby holds. However, this assumption does not hold in general social network microtext, as people often directly use informal words without any explanations or definitions.

While seminal, Li and Yarowsky's method has other shortcomings. Relying on a search engine, the system recovers only highly frequent and conventional informal words that have been defined on the web, relying heavily on the quality of Baidu's index. 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).

Normalizing informal words is another focus area in related work. An important channel for informal–formal mapping (as we review in detail later) is phonetic substitution. In work on Chinese, this is often done by measuring the Pinyin similarity [2] 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, Xia et al. (2008) propose a source channel model that is extended with phonetic mapping rules. They evaluated the model on manually-annotated phonetically similar informal–formal pairs. The disadvantage is that these rules need to be manually created and tuned. For example, $Sim(chi, qi)$ is calculated as $Sim(ch, q) * Sim(i, i)$ (here, "ch" and "q" are *Pinyin initials* and "i" is a *Pinyin final*, as per convention), in which $Sim(ch, q) = 0.8$ and $Sim(i, i) = 1.0$ are defined manually by the annotators. As informal words and their usage in microtext continually evolve, they noted that it is difficult for annotators to accurately weigh the similarities for all pronunciation pairs. We concur that the labor of manually tuning weights is unnecessary, given annotated informal–formal pairs. Finally, we make the key observation that the similarity of initial and final pairs are not independent, but may vary contextually. As such, a decomposition of $Sim(chi, qi)$ as $Sim(ch, q) * Sim(i, i)$ may not be wholly accurate.

To tackle these problems as a whole, we propose a two-step solution to the normalization task, which involves formal candidate generation followed by candidate classification. Our pipeline relaxes the strong assumptions described by prior work and achieves significant improvement over the previous state-of-the-art.

3 Data Analysis

To bootstrap our work, we analyzed sample Chinese microtext, hoping to gain insight on how informal words relate to their formal counterparts. To do this, we first needed to compile a corpus of microtext and annotate them.

We utilized the Chinese social media archive, PrEV (Cui et al., 2012), to obtain Chinese microblog posts from the public timeline of Sina Weibo[3], the most popular Chinese microtext site with over half a billion users. To assemble a corpus for annotation, we first followed the convention from (Wang et al., 2012) to preprocess and label *URLs*, *emoticons*, *"@usernames"* and *Hashtags* as pre-defined words. We then employed

[1] www.baidu.com

[2] Pinyin is the official phonetic system for transcribing the sound of Chinese characters into Latin script. $PYSim(x, y)$ is used to denote the similarity between two Pinyin string "x" and "y" hereafter.

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

Zhubajie[4], one of China's largest crowdsourcing platforms to obtain third-party (i.e., not by the original author of the microtext) annotations for any informal words, as well as their normalization, sentiment and motivation for its use (Wang et al., 2010). Our coarse-grained sentiment annotations use the three categories of "positive", "neutral" and "negative". Motivation is likewise annotated with the seven categories listed in Table 1:

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%

Table 1: Categories used for motivation annotation, shown with their observed distribution.

In total, we spent US$110 to annotate a subset of $5,500$ posts ($12,446$ sentences), in which $1,658$ unique informal words were annotated. Each post was annotated by three annotators where conflicts were resolved by simple majority. Annotations were completed after a five-week span and are publicly available[5] for comparative study.

3.1 Data Feature Analysis

From our observation of the annotated informal–formal word pairs, we identified three key channels through which the majority of informal words originate, summarized in Table 2. Here, the first column describes these channels, giving each channel's observed frequency distribution as a percentage. Together, they account for about 94% of the channels by which informal words originate. The final "Motivation (%)" column also gives the distributional breakdown of motivations behind each of the channels as annotated by our crowdsourced annotators. We now discuss each channel.

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) in the Web domain and for 99% in Chinese chats (Xia et al., 2006). In our study of the microtext domain, we found it to be responsible for 63% (Table 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" $\rightarrow$ 'b"+ "s" $\rightarrow$ "bi" + "shi" (鄙视(**bi shi**); "to despise")), each of which maps to a single Chinese character. As such, we view this special case as also following the character–character mapping.

We found that phonetic subsitutions 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), we also agree that Chinese abbreviation expansion can be modeled as **character–word** mapping. The statistics in Table 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. We 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

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

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

Channel (%)	Informal Word	Formal Word	Translation	Sentiment	Motivation (%)
Phonetic Substitutions (63)	河蟹(he2 **xie4**)	和谐(he2 **xie2**)	harmonious	positive	**sensitive** (28.9)
	鸭梨(ya1 **li2**)	压力(ya1 **li4**)	pressure	neutral	**humorous** (45.2)
	bs	鄙视(**bi shi**)	despise	negative	**euphemism** (18.7)
	乘早(**cheng2** zao3)	趁早(**chen4** zao3)	as soon as possible	neutral	**others** (7.2)
Abbreviation (19)	桌游	桌面␣游戏	board game	neutral	**terse**
	剧透	剧情␣透露	tell the spoilers	neutral	**terse** (100)
Paraphrase (12)	给力	很棒	awesome	positive	**exaggerate** (66.3)
	暴汗	非常␣尴尬	very embarrassed	negative	**terse** (27.3)
	卖萌	可爱	cute	positive	others (6.4)

Table 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.

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 2, "给力" as a whole comes from the paraphrase of the single formal word "很棒", sharing the meaning of "awesome". As another example, "暴汗" ("very embarrassed") originates from two sources: "暴" meaning "十分" ("very") and "汗" meaning "尴尬" ("embarrassed"). From this observation, we feel that both **character–word** and **word–word** mappings may adequately model the normalization process for this channel.

4 Methodology

Drawing on our observations, we propose a two step generation-classification model for informal word normalization. We 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 our identified three 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 as (桌面 or 桌上) 游戏) and paraphrase (给力 很棒 or 很好 or 厉害) channels, where synonymous formal words are equivalent. In the case where an informal word is explanable as a phonetic substitution, only one formal form is viable. Our classification model caters for these multiple explanations.

Figure 1 illustrates the framework of the proposed approach. Given an input Chinese microblog post, we 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, we search the Chinese portion of the Google Web1T 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, we generate feature vectors for each three-tuple, i.e., $< O, C(O), T >$[6], consisting of both rule-based and statistical features. These features are used in a supervised binary classifier to render the final yes (informal–informal pair) or no (not an appropriate formal word explanation for the given informal word) decision.

4.1 Pre-Processing

As an initial step, we can recognize informal words and segment the Chinese words in the sentence by applying joint inference based on a Factorial Conditional Random Field (FCRF) methodology(Wang and Kan, 2013). However, as our focus in this work is on the normalization task, we use the manually-annotated gold standard informal words (O) and their formal equivalents (T) provided in our annotated dataset. To derive the informal words' context $C(O)$, we use the automatically-acquired output of the preprocessing FCRF, although noisy and a source of error.

4.2 Formal Candidate Generation

Given the two-tuple $< O, C(O) >$ generated from pre-processing, we produce a set of hypotheses $|T|$ which are formal candidates corresponding to O.

[6]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 we set in this work to 3.

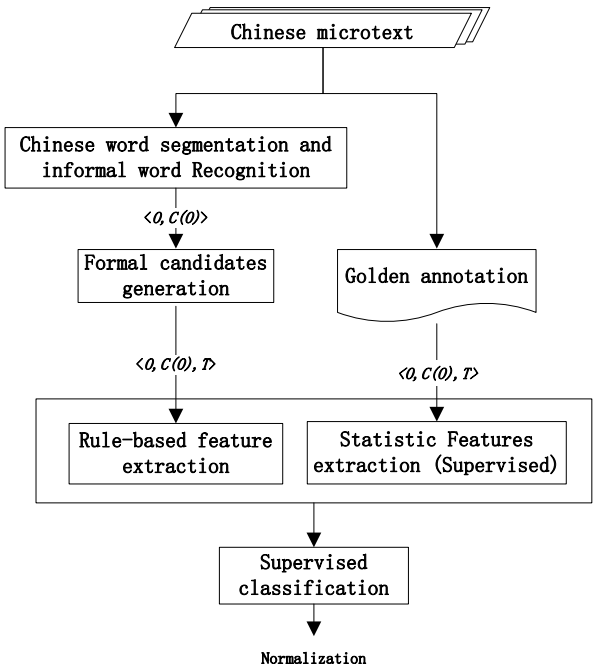

Figure 1: Our framework consists of the two steps of informal word recognition and normalization. Normalization breaks down to its component steps of candidate generation and classification.

We use two assumptions to guide us in the selection of prospective formal equivalents of O. We first discuss Assumption 1 (as [A1]):

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

To implement [A1], we define several regular expression patterns to search the Chinese Web 1T corpus, as listed in Table 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.

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

Table 3: Lexical patterns for candidate generation.

Our 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", the informal word "河蟹" is exactly equivalent to its formal equivalent "和谐", 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" and "社会".

As the Web1T 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, we 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.

We 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 our method more efficient over a brute-force approach, in exchange for loss in recall. However, we feel that this trade-off is fair: by retaining the top 1000 candidates, we 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.

4.3 Feature Extraction for Classification

For the classification step, we calculate both rule-based and statistical features for supervised machine learning. We leverage our previous observations to engineer features specific to a particular channel. We 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. We accompany each rule with an example, showing Pinyin and tones, when appropriate.

4.3.1 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 $T < c, all >$

4.3.2 Statistical Features (7 features).

We 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 " is normalized to "人ren2 才cai2", 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, we 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 our method, Pinyin similarity is formulated as:

$$PYSim(T|O) = \prod PYSim(t_i|o_i) \quad (1)$$

$$\begin{aligned} PYSim(t_i|o_i) &= PYSim(py(t_i)|py(o_i))) \\ &= \mu P(py(t_i)|py(o_i)) + \lambda P(ini(t_i)|py(o_i)) \\ &+ \eta P(fin(t_i)|py(o_i)) \end{aligned} \quad (2)$$

Here, the ti (o_i) stands for the ith 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. We use linear interpolation algorithm for smoothing, with μ, λ and η 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, we extend the source channel model (SCM) (Brown et al., 1990) to estimate the character mapping probability. In our case, SCM aims to find the formal string T that the given input O is most likely normalized to.

$$\hat{T} = \arg\max_T P(T|O) = \arg\max_T P(O|T)P(T) \quad (3)$$

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

$$P(O|T) = \prod P(o_i|t_i) \quad (4)$$

where o_i (t_i) refers to ith character of O (T). However, this SCM model suffers serious data sparsity problems, when the annotated microtext corpus is small (as in our case). To further address the sparsity, we extend the source channel model by inserting part-of-speech mapping models into Equation 4.

$$P(O|T) = \prod P'(o_i|t_i) \quad (5)$$

$$P'(o_i|t_i) = \alpha P(o_i|t_i) + \beta P(o_i|pos(t_i), pos(o_i)) \quad (6)$$

Here, let the function $pos(x)$ return the part-of-speech (POS) tag of x[7]. 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, α and β are weights for the interpolation, summing to unity.

We give the intuition for our formulation. $P(o_i|t_i)$ measures the probability of using character o_i to substitute for the given word t_i.

[7]Implemented in our system by the FudanNLP toolkithttps://code.google.com/p/fudannlp/.

$P(o_i|pos(t_i), pos(o_i))$ 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, we can optionally use formal domain synonym dictionaries to improve our model's estimation lexical and semantic similarity.

N-gram Probabilities $5\times < c, all >$. We 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: $P(W_{-1}\ T)$; $P(T\ W_1)$

5 Experiments

In our 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 we select the best classifier among three learners. The statistics reported by Li and Yarowsky (2008a) is then used as a baseline* performance. We 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 we did not re-implement their method nor execute it on our comtemporary data.

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

5.1 Data Preparation

We collected 1036 unique informal–formal word pairs with their informal contexts were collected from our annotated corpus for cross-fold validation. As any supervised classifier would do, we testing 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, we optionally employed the Cilin[8] and TYCDict[9] dictionaries.

5.2 Results

We 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.

5.2.1 Classifier choice

Table 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 training corpus. As the DT classifier performs best, we only report DT results for subsequent experiments.

Classifier	Pre	Rec	F_1
SVM	.646	.273	.383
LR	.567	.340	.430
DT (C4.5)	.886	.443	.590

Table 4: Performance comparison using different classifiers.

5.2.2 Comparison with Baseline*

To make a direct comparison with the baseline*, we perform cross-fold validation using data each of three channels separately. Since Li and Yarowsky (2008a) formalized the task as a ranking problem, we show the reported Top1 and Top10 precision in Table 5[10].

Our model achieves high precision for each channel, compared with the baseline* performance. From Table 5 we observe that normalizing words due to Phonetic Substitution is relatively easy as compared to the 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

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

[9] http://www.datatang.com/data/29207/

[10] Due to the difference in classification scheme, we recomputed the reported value, given our classification.

the estimated lexicon or semantic similarity. The low recall for the Paraphrase channel suggests the difficulty of inferring the semantic similarity between word pairs.

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	—	—	—

Table 5: Performance, analyzed per channel. "—" indicate no comparable prior reported results.

5.2.3 Final Loss Rate

We note that there is a tradeoff between the data scale and performance. By keeping the Top 1000 candidates, we 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%.

5.2.4 Channel Knowledge and Use of Formal Synonym Dictionaries

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

To investigate the impact of the formal synonym dictionaries, two configurations – with and without features derived from synonym dictionaries – were also tested. To upper bound achievable performance, we trained an oracular model with the correct channel as an input feature. In the results presented in Table 6, we see that the introduction of the features from the formal synonym dictionaries enhances performance (especially 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.

Feature set	Pre	Rec	F_1
w/o	.886	.443	.590
w	**.895**	**.583**	**.706**
w + channel	.915	.638	.752

Table 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.

5.2.5 Formal Domain Synonym Acquisition

To evaluate our method in the formal text domain, we take the synonym pairs from TYCDict as the test corpus and use the microtext data together with Cilin dictionaries as training. The experiment follows the same workflow as is done for the earlier microtext experiments, except that the context is extracted from the Chinese Wikipedia[11]. As we obtained solid performance, ($Pre = .949$, $Rec = .554$ and $F_1 = .699$), we feel that our method can be applied to synonym acquisition task in the formal domain.

6 Conclusion

Based on our observations from a crowdsourced annotated corpus of informal Chinese words, we perform a systematic analysis about how informal words originate. We show that 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 we suggest a two-stage candidate generation-classification method. The results obtained are promising, bettering the current state of the art with respect to both F_1 and loss rate. In our detailed analysis, we find that channel knowledge can still improve performance and is a possible field for future work.

References

Aditya Bhargava and Grzegorz Kondrak. 2011. 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.

[11] http://en.wikipedia.org/wiki/Wikipedia:Database_download

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

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

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

A. Cui, L. Yang, D. Hou, M.Y. Kan, Y. Liu, M. Zhang, and S. Ma. 2012. PrEV: Preservation Explorer and Vault for Web 2.0 User-Generated Content. *Theory and Practice of Digital Libraries*, pages 101–112.

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.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, pages 10–18.

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.

Z. Li and D. Yarowsky. 2008a. 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.

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

Serguei Pakhomov. 2002. 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.

Youngja Park and Roy J Byrd. 2001. 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.

Adam Pauls and Dan Klein. 2011. 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.

Aobo Wang and Min-Yen Kan. 2013. 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.

A. Wang, C.D.V. Hoang, and M.Y. 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.

K.F. Wong and Y. Xia. 2008. Normalization of Chinese Chat Language. *Language Resources and Evaluation*, pages 219–242.

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

Y. Xia, K.F. Wong, and W. Gao. 2005. NIL Is Not Nothing: Recognition of Chinese Network Informal Language Expressions. In *4th SIGHAN Workshop on Chinese Language Processing*, volume 5.

Yunqing Xia, Kam-Fai Wong, and Wenjie Li. 2006. 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.

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