Mining and Modeling Relations between Formal and Informal Chinese Phrases from Web Corpora

Zhifei Li and **David Yarowsky**
Department of Computer Science and Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218, USA
zhifei.work@gmail.com and yarowsky@cs.jhu.edu

Abstract

We present a novel method for discovering and modeling the relationship between informal Chinese expressions (including colloquialisms and instant-messaging slang) and their formal equivalents. Specifically, we proposed a bootstrapping procedure to identify a list of candidate informal phrases in web corpora. Given an informal phrase, we retrieve contextual instances from the web using a search engine, generate hypotheses of formal equivalents via this data, and rank the hypotheses using a conditional log-linear model. In the log-linear model, we incorporate as feature functions both rule-based intuitions and *data co-occurrence* phenomena (either as an explicit or indirect definition, or through formal/informal usages occurring in free variation in a discourse). We test our system on manually collected test examples, and find that the (*formal-informal*) relationship discovery and extraction process using our method achieves an average 1-best precision of 62%. Given the ubiquity of informal conversational style on the internet, this work has clear applications for text normalization in text-processing systems including machine translation aspiring to broad coverage.

1 Introduction

Informal text (e.g., newsgroups, online chat, blogs, etc.) is the majority of all text appearing on the Internet. Informal text tends to have very different style from formal text (e.g., newswire, magazine, etc.). In particular, they are different in vocabulary, syntactic structure, semantic interpretation, discourse structure, and so on. On the other hand, certain relations exist between the informal and formal text, and informal text often has a viable formal equivalent. Table 1 shows several naturally occurring examples of informal expressions in Chinese, and Table 2 provides a more detailed inventory and characterization of this phenomena[1]. The first example of informal phrase "88" is used very often in Chinese on-line chat when a person wants to say "bye-bye" to the other person. This can be explained as follows. In Chinese, the standard equivalent to "bye-bye" is "拜拜" whose *PinYin* is "BaiBai". Coincidentally, the *PinYin* of "88" is "BaBa". Because "BaBa" and "BaiBai" are near homophones, people often use "88" to represent "拜拜", either for input convenience or just for fun. The other relations in Table 1 are formed due to similar processes as will be described later.

Formal	**Informal**
拜拜 (BaiBai)[bye-bye]	88 (BaBa)
喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]
哥哥 (GeGe)[elder brother]	GG
歌迷 (GeMi)[fans]	粉丝 (FenSi)[a food]

Table 1: Example Chinese *Formal-informal* Relations. The *PinYin* pronunciation is in parentheses and an optional literal gloss is in brackets.

Due to the often substantial divergence between

[1] For clarity, we represent Chinese words in the format: Chinese characters (optional *PinYin* equivalent in parentheses and optional English gloss in brackets).

informal and formal text, a text-processing system trained on formal text does not typically work well on informal genres. For example, in a machine translation system (Koehn et al., 2007), if the bilingual training data does not contain the word "稀饭" (the second example in Table 1), it leaves the word untranslated. On the other hand, if the word "稀饭" does appear in the training data but it has only a translation "gruel" as that is the meaning in the formal text, the translation system may wrongly translate "稀饭" into "gruel" for the *informal* text where the word "稀饭" is more likely to mean "like". Therefore, as a text-normalization step, it is desirable to transform the informal text into its standard formal equivalent before feeding it into a general-purpose text-processing system. Unfortunately, there are many processes for generating informal expressions in common use today. Such transformations are highly flexible/diverse, and new phrases are invented on the Internet every day due to major news events, popular movies, TV shows, radio talks, political activities, and so on. Therefore, it is of great interest to have a *data-driven* method that can *automatically* find the relations between informal and formal expressions.

In this paper, we present a novel method for discovering and modeling the relationship between informal Chinese expressions found in web corpora and their formal equivalents. Specifically, we implement a bootstrapping procedure to identify a list of candidate informal phrases. Given an individual informal phrase, we retrieve contextual instances from the web using a search engine (in this case, *www.baidu.com*), generate hypotheses of formal equivalents via this data, and rank the hypotheses using a conditional log-linear model. In the log-linear model, we incorporate as feature functions both rule-based intuitions and *data co-occurrence* phenomena (either as an explicit or indirect definition, or through formal/informal usages occurring in free variation in a discourse). We test our system on manually collected test examples[2], and find that the (*formal-informal*) relationship discovery and extraction process using our method achieves an average precision of more than 60%. This work has applications for text normalization in many general-purpose text-processing tasks, e.g., machine translation.

To the best of our knowledge, our work is the first published machine-learning approach to productively model the broad types of relationships between informal and formal expressions in Chinese using web corpora.

2 Formal to Informal: Phenomena and Examples

In this section, we describe the phenomena and provide examples of the relations between formal and informal expressions in Chinese (we refer to the relation as *formal-informal* phrases hereafter, even in the case of single-word expressions). We manually collected 908 *formal-informal* relations, and classified these relations into four categories. We collected these pairs by investigating multiple webpages where the *formal-informal* relations are manually compiled, and then merged these seed relations and removed duplicates. In this way, the 908 examples should give good coverage on the typical categories in the *formal-informal* relations. Also, the distribution of the categories found in the 908 examples should be representative of the actual distribution of the *formal-informal* relations occurring in the real text. Table 2 presents these categories and examples in each category. In the last column, the table also shows the relative frequency of each category, computed based on the 908 examples. Recall that we represent Chinese words in the format: Chinese characters (optional *PinYin* equivalent in parentheses and optional English gloss in brackets).

2.1 Homophone

In general, a homophone is a word that is pronounced the same as another word but differs in meaning and/or written-form. Here, we use the word "homophone" in a loose way. In particular, we refer an informal phrase as a homophone of a formal phrase if its pronunciation is the same or *similar* to the formal phrase. In the three examples belonging to the homophone category in Table 2, the first example is a true homophone, while the other two are loose homophones. The third example represents a major sub-class where the informal phrase is a *number* (e.g., 88).

[2]The training and test examples are freely available at http://www.cs.jhu.edu/∼zfli.

Category	Formal	Informal	%
Homophone	版主 (BanZhu) [system administrator]	斑竹 (BanZhu) [bamboo]	4.2
	喜欢 (XiHuan)[like]	稀饭 (XiFan)[gruel]	4.4
	拜拜 (BaiBai)[bye-bye]	88 (BaBa)	21
Abbreviation	美国军队 (MeiGuoJunDui)[american army]	美军 (MeiJun)[american army]	3.8
Acronym	哥哥 (GeGe)[elder brother]	GG	12.3
	女朋友 (N*ü*PengYou)[girl friend]	GF	7.2
Transliteration	歌迷 (GeMi)[fans]	粉丝 (FenSi)[a Chinese food]	2.3
	谢谢 (XieXie)[thank you]	3Q (SanQiu)	
Others	希拉里粉丝 (XiLaLiFenSi)[fans of Hilary]	稀饭 (XiFan)[gruel]	44.8
	奥巴马粉丝 (AoBaMaFenSi)[fans of Obama]	藕粉 (OuFen)[a food]	
	超强 (ChaoQiang)[super strong]	走召弓虽 (ZouZhaoGongXu)	

Table 2: Chinese *Formal-informal* Relations: Categories and Examples. Literal glosses in brackets.

For illustrative purposes, we can present the *transformation path* showing how the informal phrase is obtained from the formal phrase. In particular, the *transformation path* for this category is "Formal → PinYin → Informal (similar or same PinYin as the formal phrase)".

2.2 Abbreviation and Acronym

A Chinese *abbreviation* of a formal phrase is obtained by selecting *one or more* characters from this formal phrase, and the selected characters can be at *any* position in the formal phrase (Li and Yarowsky, 2008; Lee, 2005; Yin, 1999). In comparison, an *acronym* is a special form of abbreviation, where only the first character of each word in the formal phrase is selected to form the informal phrase. Table 2 presents three examples belonging to this category. While the first example is an abbreviation, and the other two examples are acronyms.

The *transformation path* for the second example is "Formal → PinYin → Acronym", and the *transformation path* for the third example is "Formal → English → Acronym". Clearly, they differ in whether *PinYin* or English is used as a bridge.

2.3 Transliteration

A transliteration is transcribing a word or text written in one writing system into another writing system. Table 2 presents examples belonging to this category. In the first example, the Chinese informal phrase "粉丝 (FenSi)[a Chinese food]" can be thought as a transliteration of the English phase "fans" as the pronunciation of "fans" is quite similar to the *PinYin* "FenSi".

The *transformation path* for this category is "Formal → English → Chinese Transliteration".

2.4 Others

Due to the inherently informal and flexible nature of expressions in informal genre, the formation of an informal phrase can be very complex or ad-hoc. For example, an informal phrase can be generated by applying the above transformation rules *jointly*. More importantly, many relations cannot be described using a simple set of rules. Table 2 presents three such examples, where the first two examples are generated by applying rules *jointly* and the third example is created by decomposing the Chinese characters in the formal form. The statistics collected from the 904 examples tells us that about 45% of the relations belonging to this category. This motivates us to use a *data-driven* method to automatically discover the relations between informal and formal phrases.

3 Data Co-occurrence

In natural language, related words tend to appear together (i.e., co-occurrence). For example, *Bill Gates*

tends to appear together with *Microsoft* more often than expected by chance. Such co-occurrence may imply the existence of a relationship, and is exploited in *formal-informal* relation discovery under different conditions.

3.1 Data Co-occurrence in Definitions

In general, for many informal phrases in popular use, there is likely to be an explicit definition somewhere that provides or paraphrases its meaning for an unfamiliar audience. People have created dedicated definition web-pages to explain the relations between formal and informal phrases. For example, the first example in Table 3 is commonly explained in many dedicated definition web-pages on the Internet. On the other hand, in some formal text (e.g., research papers), people tend to define the informal phrase before it is used frequently in the later part of the text. The second example of Table 3 illustrates this phenomena. Clearly, the definition text normally contains salient patterns. For example, the first example follows the "*informal*是*formal*的意思" definition pattern, while the second example follows the pattern "*formal* (*informal*)". This gives us a reliable way to seed and bootstrap a list of informal phrases as will be discussed in Section 4.1.

Relation	**Definition Text**
(女朋友, GF)	GF是女朋友的意思。
(世界卫生组织, 世卫)	香港的食水采用世界卫生组织 (世卫)饮用水水质指引……

Table 3: Data Co-occurrence in Definitions

3.2 Data Co-occurrence in Online Chat

Informal phrases appear in online chat very often for input convenience or just for fun. Since different people may have different ways or traditions to express semantically-equivalent phrases, one may find many nearby *data co-occurrence* examples in chat text. For example, in Table 4, after a series of message exchanges, person A wants to end the conversation and types "拜拜" (meaning "bye-bye"), person B later includes the same semantic content, but in a different (more or less formal) expression (e.g. "88").

	⋮
Person A:	对不起，我要先下线了
Person A:	拜拜
Person B:	88

Table 4: Data Co-occurrence in Online Chat for Relation (拜拜, 88) meaning "bye-bye"

3.3 Data Co-occurrence in News Articles

For some *formal-informal* relations, since both of the informal and formal phrases have been used in public very often and people are normally aware of these relations, an author may use the informal and formal phrases interchangeably without bothering to explain the relations. This is particularly true in news articles for some well-known relations. Table 5 shows an example, where the abbreviation "冬奥会" (meaning "winter olympics") appears in the title and its full-form "冬季奥运会" appears in the text of the same document. In general, the relative distance between an informal phrase and its formal phrase varies. For example, they may appear in the same sentence, or in neighboring sentences.

Title	都灵**冬奥会**开幕式将激情上演
Text	新华社都灵2月9日电(记者丁莹阎涛)第20届**冬季奥运会**的开幕式将于当地时间10日晚8点在都灵奥林匹克体育场正式揭开神秘的面纱。

Table 5: Data Co-occurrence in News Article for Relation (冬季奥运会, 冬奥会) meaning "winter olympics"

4 Mining Relations between Informal and Formal Phrases from Web

In this section, we describe an approach that automatically discovers the relation between a formal phrase and an informal phrase from web corpora. Specifically, we propose a bootstrapping procedure to identify a list of candidate informal phrases. Given a target informal phrase, we retrieve a large set of of instances in context from the Web, generate candidate hypotheses (i.e, candidate formal phrases) from the data, and rank the hypotheses by using a conditional log-linear model. The log-linear model is very flexible to incorporate both the rule- and data-

driven intuitions (described in Sections 2 and 3, respectively) into the model as feature functions.

4.1 Identifying Informal Phrases

Before finding the formal phrase corresponding to an informal phrase, we first need to identify informal phrases of interest. For example, one can collect informal phrases manually. However, this is too expensive as new relations between informal and formal phrases emerge every day on the Internet. Alternatively, one can employ a large amount of formal text (e.g., newswire) and informal text (e.g., Internet blogs) to derive such a list as follows. Specifically, from the informal corpus we can extract those phrases whose frequency in the informal corpus is significantly different from that in the formal corpus. However, such a list may be quite noisy, i.e., many of them are not informal phrases at all.

An alternative approach to extracting the informal phrases is to use a bootstrapping algorithm (e.g., Yarowsky (1995)). Specifically, we first manually collect a small set of example relations. Then, using these relations as a *seed set*, we extract the text *patterns* (e.g., the definition pattern showing how the informal and formal phrases co-occur in the data as discussed in Section 3.1). With these patterns, we identify many more new relations from the data and augment them into the seed set. The procedure iterates. Using such an approach, we should be able to extract a large list of *formal-informal* relations. Clearly, the list extracted in this way may be quite noisy, and thus it is important to exploit both the data- and rule-driven intuitions to rank these relations properly.

4.2 Retrieving Data from Web

Given an informal phrase, we retrieve training data from the web on the fly. Specifically, we first use a search engine to identify a set of hyper-links that point to web pages containing contexts relevant to the informal phrase, and then follow the hyper-links to download the web pages. The input to the search engine is a text query. One can simply use the informal phrase as a query. However, this may lead to a set of pages that have nothing to do with the informal phrase. For example, if we search the informal phrase "88" (the third example in Table 2) using the well-known Chinese search engine *www.baidu.com*, none of the top-10 pages are related to the informal phrase "88". To avoid this situation, one can use a search engine that is dedicated to informal text search (e.g., *blogsearch.baidu.com*). Alternatively, one can use the general-purpose search engine but expanding the query with domain information. For example, for the informal phrase "88", we can use a query "88 网络语言", where "网络语言" means *internet language*.

4.3 Generating Candidate Hypotheses

Given an informal phrase, we generate a set of hypotheses which are candidate formal phrases corresponding to the informal phrase. We considered two general approaches to the generation of hypotheses.

Rule-driven Hypothesis Generation: One can use the rules described in Section 2 to generate a set of hypotheses. However, with this approach, one may generate an exponential number of hypotheses. For example, assuming the number of English words starting with a given letter is O($|V|$), we can generate O($|V|^n$) hypotheses given an *acronym* containing n letters. Another problem with this approach is that a relation between an informal phrase and a formal phrase may not be explained by a specific rule. In fact, as shown in the last row of Table 2, such relations consist of 44.8% of all corpus instances.

Data-driven Hypothesis Generation: With data retrieved from the Web, we can generate hypotheses by enumerating the frequent n-grams co-occurring with the informal phrase within certain distance. This exploits the *data co-occurrence* phenomena described in Section 3, that is, the formal phrase tends to co-occur with the informal phrase nearby in the data, for the multiple reasons described above. This can deal with the cases where the relation between an informal phrase and a formal phrase cannot be explained by a rule. However, it also suffers from the over-generation problem as in the rule-driven approach.

In this paper, we use the data-driven method to generate hypotheses, and rank the hypotheses using a conditional log-linear model that incorporates both the rule and data intuitions as feature functions.

4.4 Ranking Hypotheses: Conditional Log-linear Model

Log-linear models are known for flexible incorporation of features into the model. Each feature function reflects a hint/intuition that can be used to rank the hypotheses. In this subsection, we develop a conditional log-linear model that incorporates both the rule and data intuitions as feature functions.

4.4.1 Conditional Log-linear Model

Given an informal phrase (say x) and a candidate formal phrase (say y), the model assigns the pair a score (say $s(x, y)$), which will be used to rank the hypothesis y. The score $s(x, y)$ is a linear combination of the feature scores (say $\Phi_i(x, y)$) over a set of feature functions indexed by i. Formally,

$$s(x, y) = \sum_{i=1}^{K} \Phi_i(x, y) \times \alpha_i \quad (1)$$

where K is the number of feature functions defined and α_i is the weight assigned to the i-th feature function (i.e., Φ_i). To learn the weight vector $\vec{\alpha}$, we first define a probability measure,

$$P_{\vec{\alpha}}(y|x) = \frac{1}{Z(x, \vec{\alpha})} e^{s(x,y)} \quad (2)$$

where $Z(x, \vec{\alpha})$ is a normalization constant. Now, we define the regularized log-likelihood (LL_R) of the training data (i.e, a set of pairs of (x, y)), as follows,

$$LL_R(\vec{\alpha}) = \sum_{j=1}^{N} \log P_{\vec{\alpha}}(y_j|x_j) - \frac{||\vec{\alpha}||^2}{2\sigma^2} \quad (3)$$

where N is the number of training examples, and the regularization term $\frac{||\vec{\alpha}||^2}{2\sigma^2}$ is a Gaussian prior with a variance σ^2 (Roark et al., 2007). The optimal weight vector $\vec{\alpha}^*$ is obtained by maximizing the regularized log-likelihood (LL_R), that is,

$$\vec{\alpha}^* = \arg\max_{\vec{\alpha}} LL_R(\vec{\alpha}) \quad (4)$$

To maximize the above function, we use a limited-memory variable method (Benson and More, 2002) that is implemented in the TAO package (Benson et al., 2002) and has been shown to be very effective in various natural language processing tasks (Malouf, 2002).

During test time, the following **decision rule** is normally used to predict the optimal formal phrase y^* for a given informal phrase x,

$$y^* = \arg\max_{y} s(x, y). \quad (5)$$

4.4.2 Feature Functions

As mentioned before, we incorporate both the rule- and data-driven intuitions as feature functions in the log-linear model.

Rule-driven feature functions: Clearly, if a pair (x, y) matches the rule patterns described in Table 2, the pair has a high possibility to be a true *formal-informal* relation. To reflect this intuition, we develop several feature functions as follows.

- LD-PinYin(x, y): the *Levenshtein distance* on PinYin of x and y. The distance between two PinYin characters is weighted based on the similarity of pronunciation, for example, the weight $w(l, n)$ is smaller than the weight $w(a, z)$.
- LEN-PinYin(x, y): the difference in the number of PinYin characters between x and y.
- Is-PinYin-Acronym(x, y): is x a PinYin acronym of y? For example,
  Is-PinYin-Acronym(GG, 哥哥)=1,
  Is-PinYin-Acronym(GG, 兄弟)=0.
- Is-CN-Abbreviation(x, y): is x a Chinese abbreviation of y? For example,
  Is-CN-Abbreviation(美军,美国军队)=1,
  Is-CN-Abbreviation(美军,中国军队)=0.

Data-driven feature functions: As described in Section 3, the informal and formal phrases tends to co-occur in the data. Here, we develop several feature functions to reflect this intuition.

- n-gram co-occurrence relative frequency: we collect the n-grams that occur in the data within a window of the occurrence of the informal phrase, and compute their relative frequency as feature values. Since different orders of grams will have quite different statistics, we define 7 features in this category: 1-gram, 2-gram, 3-gram, 4-gram, 5-gram, 6to10-gram, and 11to15-gram. Note that the *order* n of a n-gram is in terms of number of Chinese characters instead of words.

- Features on a definition pattern: we have discussed definition patterns in Section 3.1. For each definition pattern, we can define a feature function saying that if the co-occurrence of x and y satisfies the definition pattern, the feature value is one, otherwise is zero.
- Features on the number of relevant web-pages: another interesting feature function can be defined as follows. For each candidate relation (x, y), we use the pair as a query to search the web, and treat the *number* of pages returned by the search engine as a feature value.[3] However, these features are quite expensive as millions of queries may need to be served.

Category	Feature	Weight
Rule-driven	LD-PinYin	0.800
	Len-PinYin	0.781
	Is-PinYin-Acronym	7.594
	Is-CN-Abbreviation	7.464
Data-driven	1-gram	14.506
	2-gram	108.193
	3-gram	82.975
	4-gram	66.872
	5-gram	42.258
	6to10-gram	21.229
	11to15-gram	0.985

Table 6: Optimal Weights in the Log-linear Model

5 Experimental Results

Recall that in Section 2 we categorize the *formal-informal* relations based on the manually collected relations. In this section, we use a subset of them for training and testing. In particular, we use 252 examples to train the log-linear model that is described in Section 4, and use 249 examples as test data to compute the precision.[4]

Table 6 shows the weights[5] learned for the various feature functions described in Section 4.4. Clearly, different feature functions get quite different weights. This is intuitive as the feature functions may differ in the scale of the feature values or in their importance in ranking the hypotheses. In fact, this shows the importance of using the log-linear model to learn the optimal weights in a principled and automatic manner, instead of manually tuning the weights in an ad-hoc way.

Tables 7-9 show the precision results for different categories as described in Section 2, using the rule-driven, data-driven, or both rule and data-driven features, respectively. In the tables, the precision corresponding to the "top-N" is computed in the following way: if the true hypothesis is among the top-N hypotheses ranked by the model, we tag the classification as correct, otherwise as wrong. Clearly, the larger the N is, the higher the precision is. Computing the top-N precision (instead of just computing the usual top-1 precision) is meaningful especially when we consider our relation extractor as an intermediate step in an end-to-end text-processing system (e.g., machine translation) since the final decision can be delayed to later stage based on more evidence. In general, our model gets quite respectably high precision for such a task (e.g., more than 60% for top-1 and more than 85% for top-100) when using both data and rule-driven features, as shown in Table 9. Moreover, the data-driven features are more helpful than the rule-driven features (e.g, 25.3% absolute improvement in 1-best precision), while the combination of these features does boost the performance of any individual feature set (e.g., 10.4% absolute improvement in 1-best precision over the case using data-driven features only).

We also carried out experiments (see Table 10) in the bootstrapping procedure described in Section 4.1. In particular, we start from a seed set having 130 relations. We identify the frequent patterns from the data retrieved from the web for these seed examples. Then, we use these patterns to identify many more new possible *formal-informal* relations. After the first iteration, we select the top 3000 pairs of relations matched by the patterns. The recall of a manually collected test set (having 750 pairs) on these 3000 pairs is around 30%, which is quite promising given the highly noisy data.

[3] Note that the number of pages relevant to a query can be easily obtained as most search engines return this number.

[4] Again, the training and test examples are freely available at http://www.cs.jhu.edu/∼zfli.

[5] Note that we do not use the features on definition patterns and on the number of relevant web pages, for efficiency.

Category		Precision (%)			
		Top-1	Top-10	Top-50	Top-100
Homophone	Same PinYin	31.6	47.4	68.4	73.7
	Similar PinYin	15.0	35.0	45.0	50.0
	Number	31.6	64.2	84.2	90.5
Abbreviation	Chinese abbreviation	11.8	35.3	41.2	41.2
Acronym	PinYin Acronym	39.3	82.1	91.1	92.9
	English Acronym	3.1	6.3	9.4	28.1
Transliteration		10.0	20.0	20.0	20.0
Average		*26.1*	*53.4*	*66.3*	*72.3*

Table 7: **Rule-driven Features only**: Precision on Chinese *Formal-informal* Relation Extraction

Category		Precision (%)			
		Top-1	Top-10	Top-50	Top-100
Homophone	Same PinYin	52.6	73.7	73.7	78.9
	Similar PinYin	45.0	65.0	75.0	75.0
	Number	66.3	86.3	94.7	96.8
Abbreviation	Chinese abbreviation	0.0	23.5	47.1	47.1
Acronym	PinYin Acronym	58.9	78.6	85.7	87.5
	English Acronym	25.0	46.9	68.6	68.8
Transliteration		50.0	50.0	50.0	50.0
Average		*51.4*	*71.1*	*81.1*	*82.7*

Table 8: **Data-driven Features only**: Precision on Chinese *Formal-informal* Relation Extraction

Category		Precision (%)			
		Top-1	Top-10	Top-50	Top-100
Homophone	Same PinYin	63.2	73.7	84.2	84.2
	Similar PinYin	40.0	60.0	70.0	80.0
	Number	81.1	91.6	95.8	96.8
Abbreviation	Chinese abbreviation	11.8	41.2	52.9	52.9
Acronym	PinYin Acronym	82.1	94.6	96.4	96.4
	English Acronym	21.9	46.9	56.3	59.4
Transliteration		20.0	40.0	50.0	50.0
Average		**61.8**	**77.1**	**83.1**	**84.7**

Table 9: **Both Data and Rule-drive Features**: Precision on Chinese *Formal-informal* Relation Extraction

Size of seed set	130
Size of candidate set	3000
Size of test set	750
Recall	**30%**

Table 10: Recall of Test Set on a Candidate Set Extracted by a *Bootstrapping* Procedure

6 Related Work

Automatically extracting the relations between full-form Chinese phrases and their abbreviations is an interesting and important task for many NLP applications (e.g., machine translation, information retrieval, etc.). Recently, Chang and Lai (2004), Lee (2005), Chang and Teng (2006), Li and Yarowsky (2008) have investigated this task. Specifically, Chang and Lai (2004) describes a hidden markov model (HMM) to model the relationship between a full-form phrase and its abbreviation, by treating the abbreviation as the *observation* and the full-form words as *states* in the model. Using a set of manually-created *full-abbreviation* relations as training data, they report experimental results on a *recognition* task (i.e., given an abbreviation, the task is to obtain its full-form, or the vice versa). Chang and Teng (2006) extends the work in Chang and Lai (2004) to automatically extract the relations between full-form phrases and their abbreviations, where both the full-form phrase and its abbreviation are not given. Clearly, the method in (Chang and Lai, 2004; Chang and Teng, 2006) is *supervised* because it requires the *full-abbreviation* relations as training data. Li and Yarowsky (2008) propose an *unsupervised* method to extract the relations between full-form phrases and their abbreviations. They exploit the *data co-occurrence* phenomena in the newswire text, as we have done in this paper. Moreover, they augment and improve a statistical machine translation by incorporating the extracted relations into the baseline translation system.

Other interesting work that addresses a similar task as ours includes the work on homophones (e.g., Lee and Chen (1997)), abbreviations with their definitions (e.g., Park and Byrd (2001)), abbreviations and acronyms in the medical domain (Pakhomov, 2002), and transliteration (e.g., (Knight and Graehl, 1998; Virga and Khudanpur, 2003; Li et al., 2004; Wu and Chang, 2007)).

While all the above work deals with the relations occurring within the *formal* text, we consider the *formal-informal* relations that occur across both formal and informal text, and we extract the relations from the web corpora, instead from just formal text. Moreover, our method is *semi-supervised* in the sense that the weights of the feature functions are tuned in a *supervised* log-linear model using a small number of seed relations while the generation and ranking of the hypotheses are *unsupervised* by exploiting the *data co-occurrence* phenomena.

7 Conclusions

In this paper, we have first presented a taxonomy of the *formal-informal* relations occurring in Chinese text. We have then proposed a novel method for discovering and modeling the relationship between informal Chinese expressions (including colloquialisms and instant-messaging slang) and their formal equivalents. Specifically, we have proposed a bootstrapping procedure to identify a list of candidate informal phrases in web corpora. Given an informal phrase, we retrieved contextual instances from the web using a search engine, generated hypotheses of formal equivalents via this data, and ranked the hypotheses using a conditional log-linear model. In the log-linear model, we incorporated as feature functions both rule-based intuitions and *data co-occurrence* phenomena (either as an explicit or indirect definition, or through formal/informal usages occurring in free variation in a discourse). We tested our system on manually collected test examples, and found that the (*formal-informal*) relationship discovery and extraction process using our method achieves an average 1-best precision of 62%. Given the ubiquity of informal conversational style on the internet, this work has clear applications for text normalization in text-processing systems including machine translation aspiring to broad coverage.

Acknowledgments

We would like to thank Yi Su, Sanjeev Khudanpur, and the anonymous reviewers for their helpful comments. This work was partially supported by the Defense Advanced Research Projects Agency's GALE program via Contract No. HR0011-06-2-0001.

References

S. J. Benson, L. C. McInnes, J. J. More, and J. Sarich. 2002. Tao users manual, Technical Report ANL/MCS-TM-242-Revision 1.4, Argonne National Laboratory.

S. J. Benson and J. J. More. 2002. A limited memory variable metric method for bound constrained minimization. preprint ANL/ACSP909-0901, Argonne National Laboratory.

Jing-Shin Chang and Yu-Tso Lai. 2004. A preliminary study on probabilistic models for Chinese abbreviations. *In Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing*, Barcelona, Spain (2004),pages 9-16.

Jing-Shin Chang and Wei-Lun Teng. 2006. Mining Atomic Chinese Abbreviation Pairs: A Probabilistic Model for Single Character Word Recovery. *In Proceedings of the 5rd SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia (2006), pages 17-24.

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Computational Linguistics,* 24(4):599-612.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan,Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constrantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *In Proceedings of ACL*, Demonstration Session, pages 177-180.

H.W.D Lee. 2005. A study of automatic expansion of Chinese abbreviations. MA Thesis, The University of Hong Kong.

Yue-Shi Lee and Hsin-Hsi Chen. 1997. Applying Repair Processing in Chinese Homophone Disambiguation. *In Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 57-63.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source channel model for machine transliteration. *In Proceedings of ACL 2004*, pages 159-166.

Zhifei Li and David Yarowsky. 2008. Unsupervised Translation Induction for Chinese Abbreviations using Monolingual Corpora. *In Proceedings of ACL 2008*, pages 425-433.

R. Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. *In Proceedings of CoNLL 2002*, pages 49-55.

Serguei Pakhomov. 2002. Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts. *In Proceedings of ACL 2002*, pages 160-167.

Youngja Park and Roy J. Byrd. 2001. Hybrid text mining for finding abbreviations and their definitions. *In Proceedings of EMNLP 2001*, pages 126-133.

Brian Roark, Murat Saraclar, and Michael Collins. 2007. Discriminative n-gram language modeling. *Computer Speech and Language*, 21(2):373-392.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of Proper Names in Cross lingual Information Retrieval. *In Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition.*

Jian-Cheng Wu and Jason S. Chang. 2007. Learning to Find English to Chinese Transliterations on the Web. *In Proceedings of EMNLP-CoNLL 2007*, pages 996-1004.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *In Proceedings of ACL 1995*, pages 189-196.

Z.P. Yin. 1999. Methodologies and principles of Chinese abbreviation formation. In Language Teaching and Study, No.2 (1999) 73-82.