Detecting Chinese Wish Messages in Social Media: An Empirical Study

George Chang
Graduate Institute of
Networking and Multimedia
Nation Taiwan University
Taipei,Taiwan
r99944002@csie.ntu.edu.tw

Han-Shen Huang
Bridgewell Incorporated
Taipei,Taiwan
hanshen@bridgewell.com

Jane Yung-jen Hsu
Department of Computer Science
and Information Engineering
Nation Taiwan University
Taipei,Taiwan
yjhsu@csie.ntu.edu.tw

Abstract

People have wishes and sometimes share their wishes in social media, hoping to receive supports or to find partners with the same wishes. By collecting and analyzing those wishes, we may find out not only the trend of common wishes, but also the needs of individuals. This paper presents a preliminary study of Chinese wish detection. We provide analysis on the data from Linkwish, which is a micro social network for wish sharing with users mainly from Taiwan, Hong Kong, and Macao. Then, we use SVM with various types of features to classify these messages as wish or not. Our experimental results show that some features can achieve average areas under precision-recall curves higher than 0.95 in 10-fold cross validation.

Introduction

As defined in the New Oxford dictionary of America, wish is "a hope for something to happen." It is human nature to dream for a better life. People make wishes to the falling stars and water fountains, but little did they do to fulfill them.

We get by with existing situations even if some of the solutions might be close by. Different from opinion and sentiment analysis, wishes are stronger in terms of feeling and motivation than simple opinions or comments. By studying wishes, we try to find out what people want and need in the moment. It might be possible to fill the gap by grouping people with common interests or helping them to find solutions and helpers. Thus, wish detectors have tremendous value in collecting business intelligence and public opinions. Businesses and service providers can learn from people's wishes to better identify potential market needs or to precisely define product features. Studying wishes may actually help people fulfilling them.

Social media is considered to be a good place to collect wishes, not only because there are real persons behind each post, but also that the intention can be directly linked back to the user who posted it.

In general, social medias like Facebook and Twitter contain less than 10% of wishes in the entries. Thus, wish discovery becomes an important topic of wish analysis. Fortunately, Linkwish, a micro social media dedicated to wish sharing, has higher wish rate with approximately 63%, which makes it a good starting point for observing wishes and building a wish detectors.

Unlike opinion mining and business intelligence are well developed these days, wish analysis is still at its infancy. Study wishes is valuable not only for discovering innermost feelings, perceptions of what people desire (Speer 1939) but also being a special genre of subjective expression (Goldberg et al. 2009). Wishes differ from most sentiment analysis and opinion mining by revealing what people really want to happen, instead of preference (Hu and Liu 2004). Ding, Liu, and Yu(2008) and Goldberg et al.(2009) also show that opinions and comments corpus do not indicate user actions. They presented an English wish detector based on finding general wish templates with SVM cluster and tested the model on politics and product review domains provided by Hu and Liu(2004). Another kind of wish detector by Goldberg et al.(2009) proposed a rule-based method base on parts of speech, wish-verbs and template finding. We share the common goal of classifying text into a unique set of target categories, but use different techniques catered to our specific task in Chinese. The feature generation technique for wish detection resembles template-based methods for English wish detection (Goldberg et al. 2009) and information extraction (Agichtein and Gravano 2000).

Table 1: Common wish sentences in Linkwish Corpus

Traditional Chinese	English Translation
心想事成	Things come true
新年快樂	Happy new year
天佑日本	God bless Japan
平安健康	Peace and health
找到另一伴	Find my Mr./ Ms. Right
我想下班	I want to get off work
我要減肥！	I want to lose weight
我要賺大錢	I want to earn more money
我想看電影	I want to watch a movie
我要中樂透頭彩	I want to win the lottery
希望明天好天氣	Hope tomorrow will be a sunny day

Linkwish Corpus

Linkwish is a Facebook and Google Schemer like social media specializing in wish making mainly for Chinese users. Unique from other social media, this platform enables its users to share and respond to wishes publicly with temporal and geographical information, such that people with common interests could meetup to fulfill their wishes.

Launched in Jan. 2011, Linkwish has more than 1,000,000 visits and more than 10,000 registered users mostly at Taiwan, Hong Kong, and Macao., out of which more than 2,300 users actually posted at least one wish or replied at least once. There are over 11,000 posted Chinese sentences in the corpus collected over a period of 1.5 years.

We obtained access to this set of nearly 11,000 potential wish sentences, that we call "Linkwish Corpus". (Table 1) shows selected samples of the Corpus. Some are far-reaching fantasies and aspirations, while others deal with daily concerns such as money and affairs. We are interested in most recurrent topics and scopes in wishes. We asked 5 annotators each to annotate a random sample of 1,100 sentences in scope and topic, providing them a set of scope and topic based on our observation. Surprisingly, the result of annotation shows that only 63% are real wishes, considering the wish-making nature of Linkwish. We found that 87% of corpus sentence focus on users themselves, as shown in Figure 1. The topics are much more diverse, as shown in Figure 2. There are 26% of the sentences labeled as "others", which means miscellaneous topics.In fact, many of them even do not mention or have a topic in the sentences. For example, 希望我做了對的決定 (hope I made the right choice) will also be classified as "others". The second largest topic category is concerning health. And the third is actually about mood and feelings, many of which turned out not to be real wishes.

In order to evaluate wish detector, we labeled all 11262 sentences in the corpus is or is not a wish. The annotators judge the sentences by the intention revealed in the words. Generally, the sentences are easy to judge and deviation rate are lower than 10%. The biggest ambiguity in labeling arises when the posting is asking for meetup. A sentence such as 有人晚上有空嗎?(Is there anyone free tonight?), does not definitively tell if the user really has an intention. After resolving labeling discrepancies, the Linkwish Corpus is ready for observation.

Wish Message Detection

The wish message detection methods in our paper are linear SVM classifiers with different features. We used various feature generators to transform messages into feature vectors, and then trained linear SVMs to classify messages.

A wish sentence usually consists of a wish template and a target, regardless of languages. For example, 我想看電影(I want to watch a movie) can be decomposed as 我想__ (I want to __) and 看電影 (watch a movie), which are the template and the target, respectively. Figure 3 illustrates the process. We manually selected some wish templates and targets based on human hunch to create the following two feature generators:

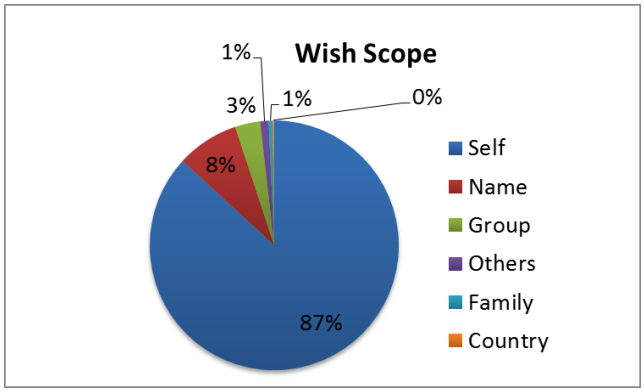

Figure 1: Wish Scope labeled by annotators on 10% random sample form dataset.

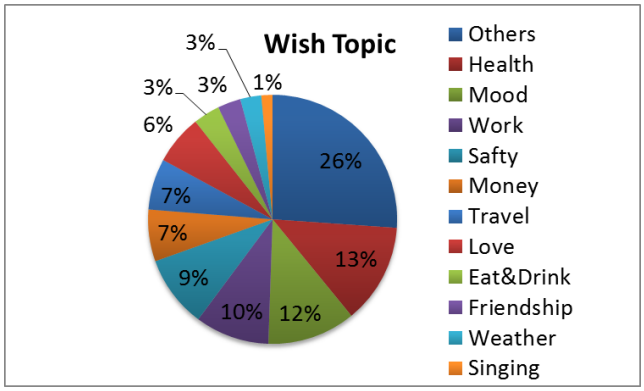

Figure 2: Wish Topic labeled by annotators on 10% random sample form dataset.

- **TEMPLATES**: We collected 18 templates in this pattern set. Here are some examples and some possible translations to English: 希望__ (wish__ or hope__), 喜歡__ (like__ or love__), 想要__ (want__), 給我__ (give me__, offer me__, or bless me__), and 許一個__ (may__ come true).
- **TARGETS**: This pattern set contains 47 targets. Most of them usually appear as positive targets that we wish to have or happen, like 幸福 (happiness), 發財 (make a lot of money), 在一起 (go steady) 升職 (get promoted) and 畢業 (graduate). We also included some negative targets that we usually wish not to have or happen, like 感冒 (catch cold), 孤單 (be lonely), and 下雨 (rain). This pattern set covers almost all targets in Linkwish corpus.

To achieve better performance, we also used four other types of feature generators that could automatically generate a variety of features for SVM to optimize the feature weights:

- ***N*-gram**: We tried 1-gram, 2-gram, and 3-gram features, where Chinese characters are the tokens. For instance, the term 在一起 (在: go; 一起: steady) is separated as (在, 一, 起) as 1-gram features, and (在一, 一起) as 2-gram features, and (在一起) as a 3-gram feature. There are 3,306 1-gram features, 56,608 2-gram features, and 114,453 3-

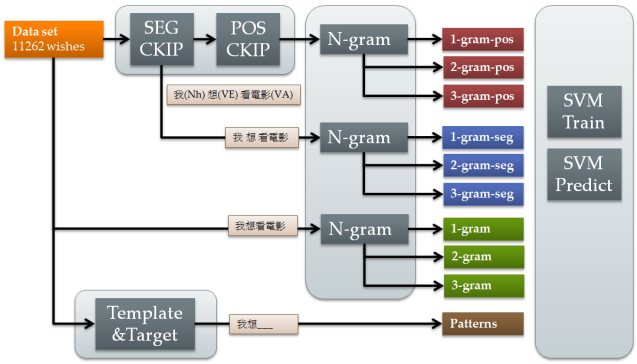

Figure 3: The feature generation flow.

gram features derived from our data set.

- **AutoTemplate**: We implemented an automatic template discovering method to determine the wish templates following the work of Goldberg et al. (2009). The idea is to separate wish sentences as template and target candidates. Then, we can obtain a directed bipartite graph, with possible templates one side, and possible targets the other side. A weighted link from a target to a template means that the combination of the target and the template really appears, where the weight is the number of observations. A reverse link from a target to a template means that the target string contains the template string, which means that the template candidate might not be a real template. We used in-degree − out-degree ≥ 3 as the threshold, and obtained 475 templates. Finally, we converted a message to a vector of indicators, showing the templates it contains.
- N**-gram-seg**: N**-gram-seg** is almost the same as N**-gram**, except that N**-gram-seg** uses relatively meaningful Chinese word segments as the token, instead of single Chinese characters. We separated each message as word segments, using the online tool provided by Chinese Knowledge and Information Processing (CKIP),Tsai and Chen (2004). The number of features in 1-gram-seg, 2-gram-seg, and 3-gram-seg are 13,863, 58,418, and 77,772, respectively.
- N**-gram-pos**: In addition to word segments, we can also obtain the part of speech of each segment by the same CKIP online tool. N**-gram-pos** means that n-gram is used, where tokens are word segments tagged with parts of speech. For example, 希望 (wish) as a noun is different from 希望 as a verb here. We have 17,129 features for 1-gram-pos, 63,836 for 2-gram-pos, and 81,351 for 3-gram-pos.

Finally, we also merged all the 10 feature generators in N-gram, AutoTemplate, N-gram-seg, and N-gram-pos as the **All** method.

Empirical Results

We used the Linkwish corpus to run 10-fold cross validation for each wish detection method described in Section . The corpus contains 11,262 instances, in which 7,144 (63.43%) are wish messages. The SVM package we used is LibSVM (2011), which can predict both the label (wish or not) and the probability to be a wish for each message. We tried some different costs in the objective function and used 0.1, which is nearly the best for all the methods. In the evaluation, we focused on the area under precision-recall curve (AUC) of each method. Since the difference among all the methods is the feature generators, we will use them to denote the corresponding message detection methods.

First, we compare **TEMPLATES** and **TARGETS**, where the features are selected manually. From Table 2 , we can see that the AUC of **TEMPLATES** is obviously greater than that of **TARGETS**, which implies that wish templates are more effective than wish targets when we use only a limited number of features. Also note that the recall of **TARGETS** is 1.0, while the precision is close to the proportion of positive examples, indicating that only few messages contains the targets we selected. As a result, SVM leaned to the majority in the training set and predicted all messages as wish ones. As for **AutoTemplate**, its AUC is also greater than that of **TARGETS**, but smaller than **TEMPLATES**.

Table 2: Empirical results of **TEMPLATES**, **TARGETS** and **AutoTemplate**

Method	Precision	Recall	AUC
TEMPLATES	0.9064	0.6247	*0.8458*
TARGETS	0.6345	1.0	0.6346
AutoTemplate	0.8021	0.6485	0.8104

Table 3 shows the results of **1-gram**, **2-gram**, and **3-gram**. By using 1-gram features, which is whether or not a particular single Chinese characters appear in a message, the AUC jumped above 0.95. **3-gram** is the best in terms of AUC, while **2-gram** also has similar performance.

Table 3: Empirical results of N**-gram** methods

Method	Precision	Recall	AUC
1-gram	0.8771	0.8932	0.9543
2-gram	0.9132	0.9167	0.9702
3-gram	0.9190	0.9163	*0.9707*

Figure 4 illustrates the precision-recall curves of **2-gram**, **1-gram**, **TEMPLATES**, **AutoTemplate**, and **TARGETS**. We omitted **3-gram** because its curve almost overlaps the curve of **2-gram**.

Now we discuss the results in Table 4. After applying word segmentation, a token in N**-gram-seg** method may contain more than one Chinese characters. Hence, **1-gram-seg** contains some features in **1-gram**, **2-gram**, and **3-gram**. We may wonder whether the features of **2-gram-seg** and **3-gram-seg** are too specific to provide enough generalization, and result in inferior performance. From Table 4, we can find that **1-gram-seg** slightly outperforms **1-gram**, while **2-gram** and **3-gram** outperform **2-gram-seg** and **3-gram-seg**. In addition, the AUC of **3-gram-seg** is even worse

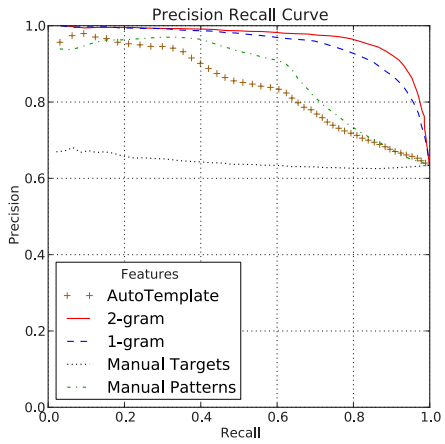

Figure 4: Precision Recall Curve of AutoTemplate, 2-gram, 1-gram, Target and Templates

than **1-gram-seg**, showing the decrease of generalization in **3-gram-seg**. We can also observe similar situation in N**-gram-pos** methods, where each word segment is associated with a part of speech.

Table 4: Empirical results of N**-gram-seg** and N**-gram-pos** methods

Method	Precision	Recall	AUC
1-gram-seg	0.8945	0.8998	0.9584
2-gram-seg	0.9158	0.9082	*0.9681*
3-gram-seg	0.7889	0.9238	0.9362
1-gram-pos	0.8944	0.9005	0.9599
2-gram-pos	0.9073	0.8988	*0.9650*
3-gram-pos	0.7748	0.9339	0.9287

Finally, we tried **All** and obtained the highest AUC in our experiments.

Table 5: Empirical results of **All**

Method	Precision	Recall	AUC
All	0.9150	0.9177	0.9711

Conclusion

In this paper, we presented our empirical study of Chinese wish detection based on Linkwish corpus. We explored a variety of wish detection methods based on SVM, and found that 1-gram methods with single Chinese characters or word segments are enough to achieve AUC higher than 0.95. We can obtain AUC as high as 0.9711 if we merge the 10 feature sets in our experiments. Since the preliminary results look promising, we would like to apply them to other copora in which users may not necessarily intent to making wishes in our future works.

Acknowledgments

The authors will likes to thanks for Linkwish.com and all of co-founders for providing their dataset and technical support.

References

Agichtein, E., and Gravano, L. 2000. Snowball:extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, 85–94.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Ding, X.; Liu, B.; and Yu, P. S. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the First ACM International Conference on Web Search and Web Data Mining (WSDM 2008)*, 231–240.

Goldberg, A. B.; Fillmore, N.; Andrzejewski, D.; Xu, Z.; Gibson, B.; and Zhu, X. 2009. May all your wishes come true: A study of wishes and how to recognize them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 263–271.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, 168–177.

Speer, G. S. 1939. Oral and written wishes of rural and city school children. 10:151—-155.

Tsai, Y.-F., and Chen, K.-J. 2004. Reliable and cost-effective pos-tagging. *International Journal of Computational Linguistics and Chinese Language Processing* 9:83–96.