An Empirical Study on Uncertainty Identification in Social Media Context

Zhongyu Wei[1], **Junwen Chen**[1], **Wei Gao**[2],
Binyang Li[1], **Lanjun Zhou**[1], **Yulan He**[3], **Kam-Fai Wong**[1]
[1]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar
[3]School of Engineering & Applied Science, Aston University, Birmingham, UK
{zywei,jwchen,byli,ljzhou,kfwong}@se.cuhk.edu.hk
wgao@qf.org.qa, y.he@cantab.net

Abstract

Uncertainty text detection is important to many social-media-based applications since more and more users utilize social media platforms (e.g., Twitter, Facebook, etc.) as information source to produce or derive interpretations based on them. However, existing uncertainty cues are ineffective in social media context because of its specific characteristics. In this paper, we propose a variant of annotation scheme for uncertainty identification and construct the first uncertainty corpus based on tweets. We then conduct experiments on the generated tweets corpus to study the effectiveness of different types of features for uncertainty text identification.

1 Introduction

Social media is not only a social network tool for people to communicate but also plays an important role as information source with more and more users searching and browsing news on it. People also utilize information from social media for developing various applications, such as earthquake warning systems (Sakaki et al., 2010) and fresh webpage discovery (Dong et al., 2010). However, due to its casual and word-of-mouth peculiarities, the quality of information in social media in terms of factuality becomes a premier concern. Chances are there for uncertain information or even rumors flooding in such a context of free form. We analyzed a tweet dataset which includes 326,747 posts (Details are given in Section 3) collected during 2011 London Riots, and result reveals that at least 18.91% of these tweets bear uncertainty characteristics[1]. Therefore, distinguishing uncertain statements from factual ones is crucial for users to synthesize social media information to produce or derive reliable interpretations, and this is expected helpful for applications like credibility analysis (Castillo et al., 2011) and rumor detection (Qazvinian et al., 2011) based on social media.

Although uncertainty has been studied theoretically for a long time as a grammatical phenomena (Seifert and Welte, 1987), the computational treatment of uncertainty is a newly emerging area of research. Szarvas et al. (2012) pointed out that "Uncertainty - in its most general sense - can be interpreted as lack of information: the receiver of the information (i.e., the hearer or the reader) cannot be certain about some pieces of information". In recent years, the identification of uncertainty in formal text, e.g., biomedical text, reviews or newswire, has attracted lots of attention (Kilicoglu and Bergler, 2008; Medlock and Briscoe, 2007; Szarvas, 2008; Light et al., 2004). However, uncertainty identification in social media context is rarely explored.

Previous research shows that uncertainty identification is domain dependent as the usage of hedge cues varies widely in different domains (Morante and Sporleder, 2012). Therefore, the employment of existing out-of-domain corpus to social media context is ineffective. Furthermore, compared to the existing uncertainty corpus, the expression of uncertainty in social media is fairly different from that in formal text in a sense that people usually raise questions or refer to external information when making uncertain statements. But, neither of the uncertainty expressions can be represented based on the existing types of uncertainty defined in the literature. Therefore, a different uncertainty classification scheme is needed in social media context.

In this paper, we propose a novel uncertainty classification scheme and construct the first uncertainty corpus based on social media data – tweets in specific here. And then we conduct experiments for uncertainty post identification and study the effectiveness of different categories of features based on the generated corpus.

[1]The preliminary study was done based on a manually defined uncertainty cue-phrase list. Tweets containing at least one hedge cue were treated as uncertain.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 58–62,
Sofia, Bulgaria, August 4-9 2013.

2 Related work

We introduce some popular uncertainty corpora and methods for uncertainty identification.

2.1 Uncertainty corpus

Several text corpora from various domains have been annotated over the past few years at different levels (e.g., expression, event, relation, sentence) with information related to uncertainty.

Sauri and Pustejovsky (2009) presented a corpus annotated with information about the factuality of events, namely *Factbank*, which is constructed based on *TimeBank*[2] containing 3,123 annotated sentences from 208 news documents with 8 different levels of uncertainty defined.

Vincze et al. (2008) constructed the BioSocpe corpus, which consists of medical and biological texts annotated for negation, uncertainty and their linguistic scope. This corpus contains 20,924 sentences.

Ganter et al. (2009) generated Wikipedia Weasels Corpus, where *Weasel tags* in Wikipedia articles is adopted readily as labels for uncertainty annotation. It contains 168,923 unique sentences with 437 weasel tags in total.

Although several uncertainty corpora exist, there is not a uniform set of standard for uncertainty annotation. Szarvas et al. (2012) normalized the annotation of the three corpora aforementioned. However, the context of these corpora is different from that of social media. Typically, these documents annotated are grammatically correct, carefully punctuated, formally structured and logically expressed.

2.2 Uncertainty identification

Previous work on uncertainty identification focused on classifying sentences into uncertain or definite categories. Existing approaches are mainly based on supervised methods (Light et al., 2004; Medlock and Briscoe, 2007; Medlock, 2008; Szarvas, 2008) using the annotated corpus with different types of features including Part-Of-Speech (POS) tags, stems, n-grams, etc..

Classification of uncertain sentences was consolidated as a task in the 2010 edition of CoNLL shared task on learning to detect hedge cues and their scope in natural language text (Farkas et al., 2010). The best system for Wikipedia data (Georgescul, 2010) employed Support Vector Machine (SVM), and the best system for biological data (Tang et al., 2010) adopted Conditional Random Fields (CRF).

In our work, we conduct an empirical study of uncertainty identification on tweets dataset and explore the effectiveness of different types of features (i.e., content-based, user-based and Twitter-specific) from social media context.

3 Uncertainty corpus for microblogs

3.1 Types of uncertainty in microblogs

Traditionally, uncertainty can be divided into two categories, namely *Epistemic* and *Hypothetical* (Kiefer, 2005). For Epistemic, there are two sub-classes *Possible* and *Probable*. For Hypothetical, there are four sub-classes including *Investigation*, *Condition*, *Doxastic* and *Dynamic*. The detail of the classification is described as below (Kiefer, 2005):

Epistemic: On the basis of our world knowledge we cannot decide at the moment whether the statement is true or false.

Hypothetical: This type of uncertainty includes four sub-classes:

- **Doxastic:** Expresses the speaker's beliefs and hypotheses.
- **Investigation:** Proposition under investigation.
- **Condition:** Proposition under condition.
- **Dynamic:** Contains deontic, dispositional, circumstantial and buletic modality.

Compared to the existing uncertainty corpora, social media authors enjoy free form of writing. In order to study the difference, we annotated a small set of 827 randomly sampled tweets according to the scheme of uncertainty types above, in which we found 65 uncertain tweets. And then, we manually identified all the possible uncertain tweets, and found 246 really uncertain ones out of these 827 tweets, which means that 181 uncertain tweets are missing based on this scheme. We have the following three salient observations:

– Firstly, there is no tweet found with the type of *Investigation*. We find people seldom use words like "examine" or "test" (indicative words of *Investigation* category) when posting tweets. Once they do this, the statement should be considered as highly certain. For example, *@dobibid I have tested the link, it is fake!*

– Secondly, people frequently raise questions about some specific topics for confirmation which expresses uncertainty. For example, *@ITVCentral*

[2] http://www.timeml.org/site/timebank/timebank.html

Can you confirm that Birmingham children's hospital has/hasn't been attacked by rioters?
– Thirdly, people tend to post message with external information (e.g., story from friends) which reveals uncertainty. For example, *Friend who works at the children's hospital in Birmingham says the riot police are protecting it*.

Based on these observations, we propose a variant of uncertainty types in social media context by eliminating the category of *Investigation* and adding the category of *Question* and *External* under *Hypothetical*, as shown in Table 3.1. Note that our proposed scheme is based on Kiefer's work (2005) which was previously extended to normalize uncertainty corpora in different genres by Szarvas et al. (2012). But we did not try these extended schema for specific genres since even the most general one (Kiefer, 2005) was proved unsuitable for social media context.

3.2 Annotation result

The dataset we annotated was collected from Twitter using Streaming API during summer riots in London during August 6-13 2011, including 326,747 tweets in total. Search criteria include hashtags like #ukriots, #londonriots, #prayforlondon, and so on. We further extracted the tweets relating to seven significant events during the riot identified by UK newspaper The Guardian from this set of tweets. We annotated all the 4,743 extracted tweets for the seven events[3].

Two annotators were trained to annotate the dataset independently. Given a collection of tweets $T = \{t_1, t_2, t_3...t_n\}$, the annotation task is to label each tweet t_i as either uncertain or certain. Uncertainty assertions are to be identified in terms of the judgements about the author's intended meaning rather than the presence of uncertain cue-phrase. For those tweets annotated as uncertain, sub-class labels are also required according to the classification indicated in Table 3.1 (i.e., multi-label is allowed).

The Kappa coefficient (Carletta, 1996) indicating inter-annotator agreement was 0.9073 for the certain/uncertain binary classification and was 0.8271 for fine-grained annotation. The conflict labels from the two annotators were resolved by a third annotator. Annotation result is displayed in Table 3.2, where 926 out of 4,743 tweets are labeled as uncertain accounting for 19.52%. *Question* is the uncertainty category with most tweets, followed by *External*. Only 21 tweets are labeled as *Dynamic* and all of them are buletic modality[4] which shares similarity with *Doxastic*. Therefore, we consider *Dynamic* together with *Domestic* in the error analysis for simplicity. During the preliminary annotation, we found that uncertainty cue-phrase is a good indicator for uncertainty tweets since tweets labeled as uncertain always contain at least one cue-phrase. Therefore, annotators are also required identify cue-phrases which trigger the sense of uncertainty in the tweet. All cue-phrases appearing more than twice are collected to form a uncertainty cue-phrase list.

Tweet#	4743	
Uncertainty#	926	
Epistemic	Possible#	16
	Probable#	129
Hypothetical	Condition#	71
	Doxastic#	48
	Dynamic#	21
	External#	208
	Question#	488

Table 2: Statistics of annotation result

4 Experiment and evaluation

We aim to identify those uncertainty tweets from tweet collection automatically based on machine learning approaches. In addition to n-gram features, we also explore the effectiveness of three categories of social media specific features including content-based, user-based and Twitter-specific ones. The description of the three categories of features is shown in Table 4. Since the length of tweet is relatively short, we therefore did not carry out stopwords removal or stemming.

Our preliminary experiments showed that combining unigrams with bigrams and trigrams gave better performance than using any one or two of these three features. Therefore, we just report the result based on the combination of them as n-gram features. Five-fold cross validation is used for evaluation. Precision, recall and F-1 score of uncertainty category are used as the metrics.

4.1 Overall performance

The overall performance of different approaches is shown in Table 4.1. We used uncertainty cue-phrase matching approach as baseline, denoted by *CP*. For *CP*, we labeled tweets containing at least one entry in uncertainty cue-phrase list (described in Section 3) as uncertain. All the other approaches are supervised methods using *SVM* based on different feature sets. *n-gram* stands for n-gram feature set, *C* means content-based feature set, *U* denotes user-based feature set, *T* represents

[3] http://www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter

[4] Proposition expresses plans, intentions or desires.

Category	Subtype	Cue Phrase	Example
Epistemic	Possible, etc.	may, etc.	It may be raining.
	Probable	likely, etc.	It is probably raining.
Hypothetical	Condition	if, etc.	If it rains, we'll stay in.
	Doxastic	believe, etc.	He believes that the Earth is flat.
	Dynamic	hope, etc.	fake picture of the london eye on fire... i hope
	External	someone said, etc.	Someone said that London zoo was attacked.
	Question	seriously?, etc.	Birmingham riots are moving to the children hospital?! seriously?

Table 1: Classification of uncertainty in social media context

Category	Name	Description
Content-based	Length	Length of the tweet
	Cue_Phrase	Whether the tweet contains a uncertainty cue
	OOV_Ratio	Ratio of words out of vocabulary
Twitter-specific	URL	Whether the tweet contains a URL
	URL_Count	Frequency of URLs in corpus
	Retweet_Count	How many times has this tweet been retweeted
	Hashtag	Whether the tweet contains a hashtag
	Hashtag_Count	Number of Hashtag in tweets
	Reply	Is the current tweet a reply tweet
	Rtweet	Is the current tweet a retweet tweet
User-based	Follower_Count	Number of follower the user owns
	List_Count	Number of list the users owns
	Friend_Count	Number of friends the user owns
	Favorites_Count	Number of favorites the user owns
	Tweet_Count	Number of tweets the user published
	Verified	Whether the user is verified

Table 3: Feature list for uncertainty classification

Approach	Precision	Recall	F-1
CP	0.3732	**0.9589**	0.5373
SVM_{n-gram}	0.7278	0.8259	0.7737
$\text{SVM}_{n-gram+C}$	0.8010	0.8260	0.8133
$\text{SVM}_{n-gram+U}$	0.7708	0.8271	0.7979
$\text{SVM}_{n-gram+T}$	0.7578	0.8266	0.7907
$\text{SVM}_{n-gram+ALL}$	**0.8162**	0.8269	**0.8215**
$\text{SVM}_{n-gram+Cue\_Phrase}$	0.7989	0.8266	0.8125
$\text{SVM}_{n-gram+Length}$	0.7372	0.8216	0.7715
$\text{SVM}_{n-gram+OOV\_Ratio}$	0.7414	0.8233	0.7802

Table 4: Result of uncertainty tweets identification

Twitter-specific feature set and *ALL* is the combination of *C*, *U* and *T*.

Table 4.1 shows that *CP* achieves the best recall but its precision is the lowest. The learning based methods with different feature sets give some similar recalls. Compared to *CP*, SVM_{n-gram} increases the F-1 score by 43.9% due to the salient improvement on precision and small drop of recall. The performance improves in terms of precision and F-1 score when the feature set is expanded by adding *C*, *U* or *T* onto *n-gram*, where *+C* brings the highest gain, and $SVM_{n-gram+ALL}$ performs best in terms of precision and F-1 score. We then study the effectiveness of the three content-based features, and result shows that the presence of uncertain cue-phrase is most indicative for uncertainty tweet identification.

4.2 Error analysis

We analyze the prediction errors based on $SVM_{n-gram+ALL}$. The distribution of errors in terms of different types of uncertainty is shown in Table 4.2. Our method performs worst on the type of *Possible* and on the combination of *Dynamic* and *Doxastic* because these two types have the least number of samples in the corpus and the classifier tends to be undertrained without enough samples.

Type	Poss.	Prob.	D.&D.	Cond.	Que.	Ext.
Total#	16	129	69	71	488	208
Error#	11	20	18	11	84	40
%	0.69	0.16	0.26	0.15	0.17	0.23

Table 5: Error distributions

5 Conclusion and future work

In this paper, we propose a variant of classification scheme for uncertainty identification in social media and construct the first uncertainty corpus based on tweets. We perform uncertainty identification experiments on the generated dataset to explore the effectiveness of different types of features. Result shows that the three categories of social media specific features can improve uncertainty identification. Furthermore, content-based features bring the highest improvement among the three and the presence of uncertain cue-phrase contributes most for content-based features.

In future, we will explore to use uncertainty identification for social media applications.

6 Acknowledgement

This work is partially supported by General Research Fund of Hong Kong (No. 417112).

References

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684.

Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. 2010. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th International Conference on World Wide Web*, pages 331–340. ACM.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the 14th Conference on Computational Natural Language Learning—Shared Task*, pages 1–12. Association for Computational Linguistics.

Viola Ganter and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009*, pages 173–176. Association for Computational Linguistics.

Maria Georgescul. 2010. A hedgehop over a max-margin framework using hedge cues. In *Proceedings of the 14th Conference on Computational Natural Language Learning—Shared Task*, pages 26–31. Association for Computational Linguistics.

Ferenc Kiefer. 2005. *Lehetoseg es szuksegszeruseg[Possibility and necessity]*. Tinta Kiado, Budapest.

H. Kilicoglu and S. Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC bioinformatics*, 9(Suppl 11):S10.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, pages 17–24.

B. Medlock and T. Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 992–999.

Ben Medlock. 2008. Exploring hedge identification in biomedical literature. *Journal of Biomedical Informatics*, 41(4):636–654.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM.

R. Saurí and J. Pustejovsky. 2009. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Stephan Seifert and Werner Welte. 1987. *A basic bibliography on negation in natural language*, volume 313. Gunter Narr Verlag.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

György Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics*.

Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan, and Shixi Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the 14th Conference on Computational Natural Language Learning—Shared Task*, pages 13–17. Association for Computational Linguistics.

V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(Suppl 11):S9.