# The UIR Uncertainty Corpus for Chinese: Annotating Chinese Microblog Corpus for Uncertainty Identification from Social Media

## Binyang Li<sup>1</sup>, Jun Xiang<sup>1</sup>, Le Chen<sup>1</sup>, Xu Han<sup>2, 7\*</sup>, Xiaoyan Yu<sup>2</sup>, Ruifeng Xu<sup>3</sup>, Tengjiao Wang<sup>4</sup>, Kam-fai Wong<sup>5, 6</sup>

School of Information Science and Technology, University of International Relations
College of Information Engineering, Capital Normal University
Harbin Institute of Technology, Shenzhen Graduate School
School of Electronics Engineering and Computer Science, Peking University
Systems Engineering and Engineering Management, The Chinese University of Hong Kong
MoE Key Lab on High Confidence Software Technologies CUHK Sub-Lab
Collaborative Innovation Center of Mathematics and Information Science, Capital Normal University
İyli, İxiang, İchen (Wuir.edu.cn; {hanxu, xyyu) (@cnu.edu.cn; xuruifeng@hit.edu.cn; kfwong@se.cuhk.edu.hk

#### **Abstract**

Uncertainty identification is an important semantic processing task, which is critical to the quality of information in terms of factuality in many NLP techniques and applications, such as question answering, information extraction, and so on. Especially in social media, the factuality becomes a primary concern, because the social media texts are usually written wildly. The lack of open uncertainty corpus for Chinese social media contexts bring limitations for many social media oriented applications. In this work, we present the first open uncertainty corpus of microblogs in Chinese, namely, the UIR Uncertainty Corpus (UUC). At current stage, we annotated 40,168 Chinese microblogs from Sina Microblog. The schema of CoNLL 2010 have been adapted, where the corpus contains annotations at each microblog level for uncertainty and 6 sub-classes with 11,071 microblogs under uncertainty. To adapt to the characteristics of social media, we identify the uncertainty based on the contextual uncertain semantics rather than the traditional cue-phrases, and the sub-class could provide more information for research on handing uncertainty in social media texts. The Kappa value indicated that our annotation results were substantially reliable.

Keywords: Chinese Microblog, Uncertainty Annotation, UIR Uncertainty Corpus

#### 1. Introduction

"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" (Szarvas et al. 2012). Many natural language processing problems are required to be solved based on the veracity of information, including question answering, information extraction, and so on, are extraordinarily need the uncertainty identification technique (Li et al., 2014).

Most participants of the CoNLL-2010 Shared Task, aiming at identifying uncertain expressions from the biological papers and Wikipedia articles, utilized linguistics features, such as lexical cues, Part-Of-Speech (POS), to detect the uncertain sentences from the given texts. (Farkas et al., 2010).

Nowadays, with the advancement of social media, there are more and more text contents containing a large amount of casual or word-of-mouth expressions. The quality of information in social media in terms of factuality has become a primary concern (Wei et al. 2013). Generally speaking, only the factual information with high credibility has the value for use. However, most expressions in social media are published with the hypothesis and episteme which cannot be decided whether it is true or false at that moment. Some previous studies investigated the existence of uncertain statements and found that they were widespread among social media, e.g., Facebook, Twitter,

Sina Weibo, etc. According to the statistic of tweet dataset (Wei et al., 2013), there were at least 18.91% of tweets possess uncertainty characteristic. As a result, the quality of information in social media has become a primary concern for many social media oriented tasks, such as rumor detection (Qazvinian et al., 2011) and credibility evaluations (Jaworski et al., 2014). Therefore, uncertainty identification is significant for users to synthesize information to derive reliable interpretation.

To do this, there are a few such corpora: the primal corpora are BioScope corpus annotating uncertainty and their scope in biological papers and the Wikipedia dataset for CoNLL 2010 shared task (Vincze, 2010).

Similar work for Chinese has been reported by Ji (2010), who constructed the Chinese corpus from newspapers for Chinese uncertainty identification. However, their data was annotated exactly based on the cue-phrases and size of corpus was a bit small for some uncertainty detection systems. In addition, Chinese uncertain cue phrases were not always indicating the uncertainty of sentences. For instance, "好像(likely)" could express uncertain with the *Possible* sub-class of uncertainty, or regard as metaphor implied certainty based on the implicit semantics.

More importantly, unlike the biological papers and Wikipedia articles, the texts in social media are usually short and informal. Due to the word count limit and casual expression, many cue phrases are expressed in substandard shape or even omitted from sentences. In this case, the uncertain semantics will be implicitly conveyed by the whole sentence rather than by the explicit cue phrases.

<sup>\*</sup> Corresponding Author

In this work, we target on annotating social media texts for uncertainty identification in Chinese language. Our corpus was collected from Sina Microblogs (also referred as Sina Weibo) during the 2010 Shanghai World Expo. After removing error and invalid sentences, 11,071 out of 40,168 microblogs were annotated as uncertain.

To summary, this paper has the following contributions:

- (1) we are the first to annotate a Chinese social media corpus, namely the UIR Uncertainty Corpus (UUC), for uncertainty identification task;
- (2) UUC corpus is constructed by 40,168 Chinese Sina Microblogs and 11,071 were labeled uncertain following our proposed scheme for Chinese with the *Kappa* value indicating substantially reliable;
- (3) UUC corpus cannot only contained certainty or uncertainty classification but also annotated six subcategories of uncertainty as *Possible*, *Condition*, *Doxastic*, *Hope*, *External*, *Question*.

The remainder of this paper is organized as follows: Section 2 gives the related work and illustrate the differences between this work and the previous work. Section 3 details the annotation scheme of uncertainty in social media texts. Section 4 presents the annotation results and some important statistics of the corpus. Section 5 concludes this work.

## 2. Related Work

Most of the current uncertainty annotation work followed the task of uncertainty identification in CoNLL 2010, in which the uncertainty texts were annotated based on the cue-phrases. It is because that the uncertainty cues commonly exist in uncertain sentences and express the . It is true that the cue words are an efficient indicator for uncertainty, but these cues are usually invalid for Chinese in social media context. To identify uncertainty clearly, our first priority is to figure out the types of uncertainty in microblogs. Kiefer (2005) pointed out that uncertainty can be divided into Epistemic and Hypothetical. Epistemic contains Possible and Probable. In fact, the two subclasses are fairly similar in Chinese. Wei (2013) observed Question and External frequently appeared on these posts or comments which reveals uncertainty. Considering Dynamic, one sub-class of Hypothetical, hardly existing in Chinese microblogs, we removed this label in our annotation scheme.

At present, there are several corpora in different domains: (1) BioScope corpus (Vincze et al., 2008) annotated uncertainty, negation sentences and their scope in biomedical texts containing 20,879 sentences from 3,236 documents

- (2) the dataset for CoNLL'2010 shared task (Vincze, 2010) consisted of biological part of BioScope corpus and the selection of Wikipedia articles, which annotated uncertain sentences and cues.
- (3) Uncertainty Corpus in complex spoken dialogue systems derived from 120 digital dialogues recording from 60 students, totaling 2,171 turns for students and 2,531 turns for tutor.
- (4) The Scientific Literature Corpus for Chinese (Chen et al., 2013) including 19 full papers annotated negation, uncertainty and their cues as well as scope.

(5) Feng Ji (2010) presented a corpus annotated uncertain cues and sentences about 50 news containing 8,000 sentences.

Although in different fields have existed uncertainty corpora, there is no standard uncertain corpus for Chinese in social media texts. The expression of uncertainty in social media is fairly different from that in formal text in a sense. Therefore, our study seeks to fill this gap by presenting the Sina Weibo corpus, which consists of 40,168 microblogs annotated for uncertainty and their subclasses.

## 3. Annotation Scheme

## 3.1 Types of Uncertainty in Chinese Microblogs

The annotation scheme adopted in this work refer to the uncertainty categories of tweets by Wei, et al. (2013). However, due to the different habits or uses of languages between English and Chinese, there are some adjustments in our annotation scheme. Generally, uncertainty of the sentences can be divided into Epistemic and Hypothetical. For Epistemic, there are two sub-classes Possible and *Probable.* Compared to the Chinese uncertainty corpora, we found that the sub-classes of *Possible* and *Probable* are very similar in Chinese uncertainty expressions. These two sub-classes are merged as one class *Possible*. In addition, we also found that of all the microglobs there was no expressions belonging to the sub-class of *Dynamic* in Chinese social media. So, we also remove the sub-class of Dynamic. On the contrary, we found that in Sina microblogs, there were a lot of sentences expressing the uncertainty statement with expected semantics.

Base on the above observations, we propose a variant of uncertainty sub-class in Chinese social media corpus by removing the *Dynamic* type and adding the sub-class of *Hope*, as shown in Figure 1.

We modify and redefine the uncertainty scheme into six sub-classes as follows:

- *Question*: Sentences are used to ask for information or to test someone's knowledge.
- External: Speaker repeat exactly what someone else has said or written.
- *Doxastic*: Expresses the speaker's beliefs and hypotheses.
- *Hope*: Expresses the speaker want something to happen or be true and to believe that it is possible or likely.
- Possible: Expresses something it can be done or achieved.
- Condition: Proposition under condition.

Examples of each sub-class of uncertain sentences are selected from the corpus:

(1) Doxastic

我相信,华夏文艺复兴正在走来。

I believe that the Chinese Renaissance is coming.

(2) Question

网游客户会对媒体品质有影响吗?

Will online visitors have an impact on media quality?

(3) External

据说那个叫苏琴的人已经当了年级组长

It is said that the person who called Suqin has been a grade leader

(4) *Hope* 

希望我们302宿舍以后可以这样,嘿嘿

Hope our 302 dormitory later can be so, heihei

(5) Condition

如果这是真的,不忍想象他们心中藏着怎样的悲痛。 If this is true, can not bear to imagine how the hearts of their hidden grief.

(6) Possible

今早一饭四菜,好像做多了

Four dishes this morning, it seems to do more

## 3.2 Annotation Process

During our annotation process, 3 annotators were trained to annotate the dataset independently. Given a collection of microblogs  $M=\{m_1, m_2, m_3, m_4, \dots m_n\}$ , the annotation task is to label each microblog  $m_1$  as either uncertain or certain.

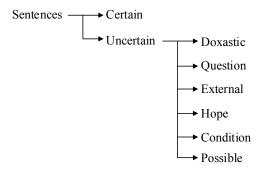


Figure 1: The annotation scheme for Chinese uncertainty annotation in social media.

Moreover, the uncertainty assertions are to be identified in terms of the judgments about author's intended meaning or the implicit semantics rather than the presence of uncertain cue-phrase. The sub-classes are also required to be further labeled according to the scheme as shown in Figure 1.

## 4. Results

The corpus we annotated was collected from Sina Weibo during the 2010 Shanghai World Expo. Currently, we have already annotated 40,168 Chinese microblogs after removing error and invalid sentences, in which we found 11,071 uncertain microblogs. This annotation work made out the UIR uncertainty corpus (UUC) in Chinese towards social media texts. Compared to the previous work on Chinese uncertainty corpus, UUC is the first Chinese corpus for uncertainty identification in social media, and the quantity of annotation is much larger than other Chinese uncertainty corpus in newspaper (Ji et al., 2010). Therefore, we can conclude some characteristics of uncertain expressions in Chinese and we further annotated six sub-class labels under uncertainty as shown in Table 1.

The Kappa coefficient (Carletta, 1996) indicating interannotator agreement was 85.68% for the certain/uncertain binary classification which means our results were substantially reliable. The disagreements in annotation work mainly come from the conflict understanding to certain ambiguous phrases, e.g., "有木有" (did you...?), "如果" (if), and we also list the specific statistics of the agreement in each sub-class, as shown in Table 2.

Refer to Table 1, there are 11,071 out of 40,168 microblogs are labeled as uncertain accounting for 27.56%.

Question is the uncertainty category with the most microblogs, followed by *Possible*. However, only 117 microblogs are labeled as *External*. We argued that there might be other uncertain sub-classes in such microblogs apart from *External*. But only can we select one more obvious label between these for these microblogs, which might replace *External*.

As illustrated in Table 2, most of the annotation results were considerably consistent. However, the *Kappa* coefficients for *External* and *Doxastic* indicated lower than the others since the amount of microblogs for *External* and *Doxastic* was smaller compared with others. Besides, our scheme is designed for single label annotation, there exists the situation that more than one label should be annotated for the same microblog.

Microblog#	40,168
Certainty#	29,097
Uncertainty#	11,071
Possible#	2,973
Condition#	1,029
Doxastic#	470
Hope#	612
External#	117
Question#	5,870

Table 1: Statistics of annotation results

Certainty/Uncertainty	0.8568
Possible#	0.9141
Condition#	0.9651
Doxastic#	0.8502
Hope#	0.8804
External#	0.8572
Question#	0.9237

Table 2: The Kappa value statistics for each sub-class

Corpus	UUC	Ji's
Total #	40,168	10,000
Uncertain#	11,071	2,858
Without Cue#	2,466	23
Percentage	22.28%	0.8%

Table 3: Comparison between UUC and Ji's corpus on cue phrases percentage

Since our argument for the annotation in social media is that the word count limit and casual expression, many cue phrases are expressed in substandard shape or even omitted from sentences. We also list the statistics between two Chinese corpus for uncertainty identification, our UUC and Ji's (2010). Note that among 2,858 uncertain sentences in the Ji's corpus, there are only 23 uncertain sentences (0.8%) not containing cue phrases. On social media dataset, however, 22.28% uncertain sentences are not containing cue phrases. So the statistics also prove our assumption.

## 5. Conclusion and Future Work

In this paper, we have presented an uncertain annotation scheme for Chinese microblogs and the first open Chinese uncertainty corpus in social media texts. 11,071 out of 40,168 Chinese Sina Microblogs were labeled uncertain following our proposed scheme with the *Kappa* value over 0.85 that indicated substantially reliable. Specifically, we

define six sub-classes for the uncertain classification which made the corpus more detailed and useful for other related NLP applications in the future.

In the future, we are planning to annotate more uncertain microblogs about the hot issue for the development of social media applications, such as rumor detection, user profile, etc.

## 6. Acknowledgements

This paper is funded by the National Natural Foundation of China 61502115, 61602326, U1636103, U1536207, 61572043, 61672361, 61632011, the Hong Kong Applied Science and Technology Research Institute Project 7050854, the Fundamental Research Fund for the Central Universityies 3262015T70, 3262017T12, and the Shenzhen Foundational Research Funding 20170307150024907. We would also like to thank Nuo Chen, Ning Jiang, Lingchen Lian, Kaimin Zhou, and Haotian Ren for their work on corpus.

## 7. References

- Castor, A. and Pollux, L. E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.
- Grandchercheur, L.B. (1983). Vers une modélisation cognitive de l'être et du néant. In S.G Paris, G.M. Olson, & H.W. Stevenson (Eds.), *Fondement des Sciences Cognitives*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 6-38.
- Strötgen, J. and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 3746–3753, Istanbul, Turkey, may. European Language Resource Association (ELRA).
- Superman, S., Batman, B., Catwoman, C., and Spiderman, S. (2000). Superheroes experiences with books. The Phantom Editors Associates, Gotham City, 20th edition.

- Vincze, V. (2016). Detecting Uncertainty Cues in Hungarian Social Media Texts.
- Wei, Z., Chen, J., Gao, W., Li, B., Zhou, L., & He, Y., et al. (2013). An Empirical Study on Uncertainty Identification in Social Media Context. Meeting of the Association for Computational Linguistics (pp.58-62).
- Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: identifying misinformation in microblogs. Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John Mcintyre Conference Centre, Edinburgh, Uk, A Meeting of Sigdat, A Special Interest Group of the ACL (pp.1589-1599). DBLP.
- Jaworski, W., Rejmund, E., & Wierzbicki, A. (2014). Credibility Microscope: Relating Web Page Credibility Evaluations to Their Textual Content. Ieee/wic/acm International Joint Conferences on Web Intelligence (Vol.1, pp.297-302). IEEE Computer Society.
- Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. Bmc Bioinformatics, 9(Suppl 11), S9-S9.
- Vincze V. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text[C]// Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2010:1-12.
- Feng Ji, Xipeng Qiu, Xuanjing Huang. (2010). Exploring uncertainty sentences in Chinese. In The 6th China Conference on Information Retrieval
- Kiefer, F. (2005). Lehetoseg es szuksegszeruseg [Possibility and necessity]. Tinta Kiadó, Budapest.
- Chen, Z., Zou, B., Zhu, Q., & Li, P. (2013). The Scientific Literature Corpus for Chinese Negation and Uncertainty Identification. Chinese Lexical Semantics. Springer Berlin Heidelberg.
- Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, cmp-lg/9602004(2), págs. 249-254.