TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 11/14 pp117–126
DOI: 10.26599/TST.2019.9010022
Volume 25, Number 1, February 2020

An Attention-Based Neural Framework for Uncertainty Identification on Social Media Texts

Xu Han, Binyang Li*, and Zhuoran Wang

Abstract: Uncertainty identification is an important semantic processing task. It is crucial to the quality of information in terms of factuality in many applications, such as topic detection and question answering. Factuality has become a premier concern especially in social media, in which texts are written informally. However, existing approaches that rely on lexical cues suffer greatly from the casual or word-of-mouth peculiarity of social media, in which the cue phrases are often expressed in substandard form or even omitted from sentences. To tackle these problems, this paper proposes an Attention-based Neural Framework for Uncertainty identification on social media texts, named ANFU. ANFU incorporates attention-based Long Short-Term Memory (LSTM) networks to represent the semantics of words and Convolutional Neural Networks (CNNs) to capture the most important semantics. Experiments were conducted on four datasets, including 2 English benchmark datasets used in the CoNLL-2010 task of uncertainty identification and 2 Chinese datasets of Weibo and Chinese news texts. Experimental results showed that our proposed ANFU approach outperformed the-state-of-the-art on all the datasets in terms of F1 measure. More importantly, 41.37% and 13.10% improvements were achieved over the baselines on English and Chinese social media datasets, respectively, showing the particular effectiveness of ANFU on social media texts.

Key words: uncertainty identification; attention; neural networks; social media

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"[1]. "*London zoo was probably attacked*" is an example of an uncertain sentence. The identification of uncertainty is significant to the trustworthiness of many Natural Language Processing (NLP) techniques and applications, such as question answering and information extraction[2].

The CoNLL-2010 Shared Task aimed at identifying uncertainty in biological papers and Wikipedia articles written in English[3,4]. Most participants utilized linguistics features, e.g., lexical cues and Parts-Of-Speech (POS), to detect the uncertain sentences from the texts.

Recently, with the growing popularity of social media, there exist more and more texts consisting of casual or word-of-mouth expressions. The quality of information in social media in terms of factuality has become a premier concern[5]. The generation and propagation of uncertain information leads to rumors flooding social media and influencing the real world. For example, the 2011 London riots were partly owing to the spread of uncertain information

• Xu Han is with the International Science and Technology Cooperation Base of Electronic System Reliability and Mathematical Interdisciplinary, Information Engineering College, Capital Normal University, Beijing 100048, China. E-mail: hanxu@cnu.edu.cn.
• Binyang Li is with the School of Information Science and Technology, University of International Relations, Beijing 100091, China. E-mail: byli@uir.edu.cn.
• Zhuoran Wang is with Tricorn (Beijing) Technology Co. Ltd, Beijing 100029, China. E-mail: wangzhuoran@trio.ai.
∗ To whom correspondence should be addressed.
Manuscript received: 2018-07-29; revised: 2018-12-31; accepted: 2019-03-04

among social media, such as Twitter and Facebook. Therefore, uncertainty identification (i.e., identifying uncertain sentences) is becoming increasingly critical to help users to synthesize information to derive reliable interpretations.

However, unlike biological papers and Wikipedia articles, texts in social media are usually short and informal. Due to the word limits and casual forms of expression, many cue phrases are expressed in a substandard shape or even omitted from sentences entirely. In this form, the uncertain semantics are implicitly conveyed by the whole sentence rather than explicitly by cue phrases. Existing approaches based on cue phrases for uncertainty identification are therefore ineffective for social media texts, and even underperform on formal texts. In the CoNLL-2010 Shared Task, the participants all achieved better results on the biological dataset than on the Wikipedia dataset, indicating that the more formal the article is, the easier it is to judge sentence uncertainty. As a result, uncertainty identification on Chinese social media texts has become a major challenge requiring more semantic information to solve.

To judge the uncertainty of Chinese text on social media based on semantics, we turned to deep learning, which can effectively express the semantics of words and sentences. Bahdanau et al.[6] applied a Recurrent Neural Network (RNN) with attention mechanism to machine translation; their model makes the semantics and relation between words in both languages clearer. Kim[7] utilized Convolutional Neural Networks (CNNs) to classify sentences and achieved good results, showing CNNs have a unique advantage both at images and text classifying tasks. Considering these studies, we decided to combine the two model structures to solve the uncertainty identification problem.

This paper proposes an Attention-based Neural Framework for the Uncertainty identification on social media texts, named ANFU. ANFU incorporates attention mechanisms into Long Short-Term Memory (LSTM) networks to represent the semantics of the context in a sentence, and uses CNNs for the uncertainty identification. Benefitting from the attention mechanisms, the key elements of sentences can be highlighted and the hidden semantics can be captured, which will enable us to detect uncertainty based on the context of the whole sentence instead of depending on the cue-phrases.

The contributions of this paper are as follows:

- We propose an attention-based neural framework (LSTM-CNN) for uncertainty identification on social media texts, which can indiscriminately focus on the words, regardless of the presence of cue-phrases, that have decisive effect on uncertain semantics, without using extra knowledge or external NLP components.
- The first annotated corpus of Chinese social media dataset is constructed for uncertainty identification, which consists of 11 071 uncertain sentences out of 30 000 sentences from a Chinese Weibo dataset.
- Experiments are conducted on the CoNLL-2010 English benchmark datasets, i.e., Wikipedia and biological datasets. On these, F1-measures of 70.02% and 87.21% were achieved with 13.10% and 2.2% improvement over the baseline, respectively.
- We also conduct experiments on Chinese Weibo and news datasets, on which F1-measures of 78.19% and 73.95% were achieved with about 41.37% and 4.8% improvement over the baseline, respectively. The experimental results attest to the effectiveness of ANFU on social media texts.

The remainder of the paper is organized as follows: Section 2 summarizes the related work, Section 3 describes our proposed methods, Section 4 presents corpus annotation as well as the experimental results, and Section 5 concludes the paper.

2 Related Work

Uncertainty identification has attracted much attention in NLP). The CoNLL-2010 Shared Task aimed at detecting uncertainty cues in English-language biological papers and Wikipedia articles[3]. Recently, a special issue of the journal *Computational Linguistics* (vol. 38, no. 2) was dedicated to detecting modality and negation in natural language texts[8]. Most of the existing approaches can be classified as rule-based[9, 10] machine learning methods, such as Medlock's research on biomedical literature[11], and the work of Fernandes et al.[12], Li et al.[13], Tang et al.[14], and Zhang et al.[15] at CoNLL2010, which mostly applied various supervised approaches to the annotated corpus to incorporate different types of linguistic features such as POS tags, word stems, *n*-grams, and so on. Velldal[16] in 2010 constructed a cue-lexicon to describe the context, which was applied into a binary classifier for detection.

The above approaches mainly focused on English-language texts, and we are aware of just one study in 2010 by Ji et al.[17] aiming at Chinese texts. In that study, a supervised method with lexical features

was proposed and evaluated on an annotated corpus consisting of Chinese news data.

Regarding uncertainty identification on social media texts, Wei et al.[5] conducted an empirical study that accounted for features beyond plain text, such as the number of tweets and their relationships, and Vincze[18, 19] proposed to use lexical, morphological, syntactic, semantic, and discourse-based features in a supervised classifier for detecting uncertainty in Hungarian social media texts.

Recently, deep learning has become popular in studies of NLP, especially for text classification. CNNs, first widely used in the field of machine vision, have been applied to NLP over recent years. Zhang and Wallace[20] conducted a sensitive analysis of one-layer CNNs, showing that the CNN model achieved state-of-the-art results in most sentence classification tasks through elaborate settings. Yang et al.[21] proposed hierarchical attention networks for document classification. Their work did not only improve the accuracy of the document classification, but also help people to understand how the attention mechanism works though the visualization of the attention weight mechanism.

Deep learning has also been used for uncertainty identification tasks. Adel and Schutze[22] presented an attention architecture for uncertainty detection, using a CNN or RNN with an attention mechanism to achieve state-of-the-art results on the CoNLL-2010 benchmark datasets. An external lexicon of seed cue words or phrases were also incorporated into the word embedding, and with this external knowledge their model performed well on English datasets. However, this model is suboptimal for social media texts, because of the long sentences frequently occurring on social media in both English and Chinese, and the use of a CNN or RNN leading to a loss of semantics.

In summary, our approach has three major differences from previous work: (1) Our model only uses word embedding, with no extra knowledge or external systems or cue words; (2) Our proposed neural networks use LSTM and the attention mechanism, which can represent well the long sentences and substandard expressions typical of social media texts to generate the semantic focus; and (3) Regarding the experimental datasets, we are the first to construct a Chinese social media corpus for evaluating uncertainty identification.

3 An Attention-Based Neural Framework for Uncertainty Identification

In this section, we will introduce our attention-based neural framework, named ANFU. Figure 1 illustrates the architecture of ANFU, which consists of three components: word representation, sentence representation, and convolutional classification. They can be summarized as follows.

- Word representation accepts an input sentence and maps each word into a k-dimensional vector of embedding.

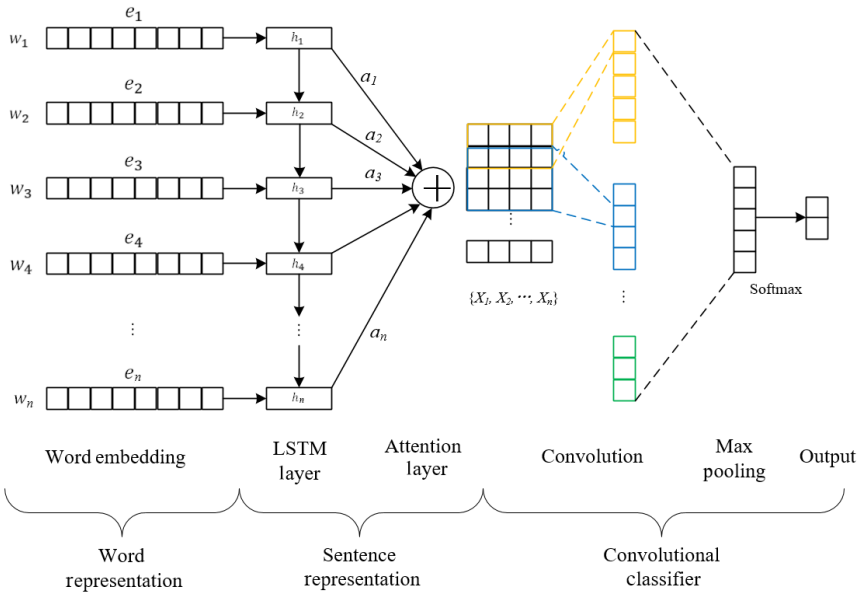


Fig. 1 Architecture of our framework.

• Sentence representation utilizes LSTM to get high-level features for more accurate word semantics representation, produces a weight vector based on attention mechanisms, and then merges word-level features by the weight vectors to highlight the words that are important for uncertainty identification.

• Convolutional classification extracts n-gram features from the sentence and selects most important part for uncertainty identification. After a full connection layer, it outputs the prediction result using a softmax function.

A more detailed description of these components is given in the following subsections.

3.1 Word representation

Suppose there is an input sentence $S = \{w_1, w_2, \ldots, w_n\}$ with n words and an embedding matrix $\boldsymbol{M} \in \mathbf{R}^{k\times|V|}$ to translate words into word vectors, where V is a fixed-sized vocabulary and k is the dimension of word embedding. For each word w_i from the input, we look up the embedding matrix to find the k-dimensional real-value vector $\boldsymbol{e}_i \in \mathbf{R}^k$ as the representation w_i.

For Chinese, there may be some words Out of Our Vocabulary (OOV). To solve this problem of rare words that cannot be represented by vectors, we discard these words. This is acceptable because OOV rarely occurs and it is very common for people when reading to skip difficult words that have little effect on understanding the whole sentence. When converting a sentence word into a vector, we set the maximum length of a sentence; sentences less than the maximum length will be padded by zero vectors.

3.2 Sentence representation

The sentence representation component consists of two layers: an LSTM layer and an attention layer.

3.2.1 LSTM networks

LSTM networks were firstly proposed to overcome the gradient vanishing problem, and an adaptive gating mechanism was introduced to decide the degree to which LSTM units keep the previous state and memorize the extracted features of the current data input[23]. Many LSTM variants have been proposed, such as Sundermeyer's improvements to language modeling[24] and the research of Yao et al.[25] into depth-gated recurrent neural networks.

We apply the variant of LSTM networks that was proposed by Graves in Ref. [26] to represent the complete semantics of a sentence. In our LSTM-based neural networks, the inputs are word vectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ and the outputs are hidden states $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\}$. There are three types of gates: one input gate *input*, one forget gate f, and one output gate o. Given the current input $\boldsymbol{e}_i$ together with the cell state c_i generated by previous cells, the combination of these gates will decide to what degree we should adopt the current input over the contents stored in memory. Our LSTM can be computed by the following equations:

$$input = \text{sigm}\{\boldsymbol{W}_i[D(\boldsymbol{h}_{i-1}), \boldsymbol{e}_i] + \boldsymbol{b}_i\} \quad (1)$$

$$f = \text{sigm}\{\boldsymbol{W}_f[D(\boldsymbol{h}_{i-1}), \boldsymbol{e}_i] + \boldsymbol{b}_f\} \quad (2)$$

$$o = \text{sigm}\{\boldsymbol{W}_o[D(\boldsymbol{h}_{i-1}), \boldsymbol{e}_i] + \boldsymbol{b}_o\} \quad (3)$$

$$g = \tanh\{\boldsymbol{W}_g[D(\boldsymbol{h}_{i-1}), \boldsymbol{e}_i] + \boldsymbol{b}_g\} \quad (4)$$

$$c_i = f \odot c_{i-1} + input \odot g \quad (5)$$

$$\boldsymbol{h}_i = o \odot \tanh c_i \quad (6)$$

where D is a dropout operation, sigm is the sigmoid function, tanh is the tanh function, $\boldsymbol{W}$ and $\boldsymbol{b}$ are the parameters that need to be learnt, $\odot$ is the elementwise multiplication.

In this way, the current cell state c_i will be generated by calculating the weighted sum of the previous cell state and the information currently generated by the cell. Because the same word may have different meanings in different contexts, only by incorporating the contextual information into the representation of the word can we express a word's meaning exactly. LSTM networks encode every word and bring previous information to bear on those words, so that each hidden state $\boldsymbol{h}_i$ can represent the meaning of a word in the specific sentence more accurately.

3.2.2 Attention

Since not all the words in a sentence contribute equally to uncertainty identification, we adopt attention mechanisms to generate better sentence representation with a semantic focus.

We calculate the attention α_i for each word w_i as follows:

$$\alpha_i = \frac{\exp(\boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}_r\boldsymbol{h}_i + \boldsymbol{b}_r))}{\sum_i \exp(\boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}_r\boldsymbol{h}_i + \boldsymbol{b}_r))} \quad (7)$$

where $\boldsymbol{v}$, $\boldsymbol{W}_r$, $\boldsymbol{b}_r$ are model parameters that need to be learnt. Unlike other attention models that sum up the product of the hidden states and their respective weights[21], we concatenate them so that all the hidden states sequences generated by word vectors can be maintained, and can then be used in the subsequent CNN component to obtain the most important features from all the words' hidden states vectors.

$$\boldsymbol{X}_{1:n} = \alpha_1 \boldsymbol{h}_1 \oplus \alpha_2 \boldsymbol{h}_2 \oplus \cdots \oplus \alpha_n \boldsymbol{h}_n \quad (8)$$

Benefitting from the attention mechanisms, the key elements in sentences can be highlighted and richer semantics can be conveyed by the encoded words (as illustrated below in our experiment). Then $\boldsymbol{X}_{1:n}$ serves as the input for the CNNs.

3.3 Convolutional classifier

CNNs are widely used and have achieved state-of-the-art results in many classification tasks. Collobert et al.[27] proposed a sentence-based network using CNNs. Inspired by their CNN architecture, we design our CNNs to determine whether a sentence is certain or uncertain. We take the sentence representation $\boldsymbol{X}_{1:n}$ carrying the hidden state of each word, as input, and return the result of the uncertainty identification, result $\in$ {certain, uncertain}, as the output.

Our CNNs involve a filter $f \in \mathbf{R}^{l \times k}$, which is applied to a window of length l sliding over the hidden states of LSTM networks to produce a new feature. For example, the i-th new feature NF_i can be computed by the following equation:

$$NF_{l_i} = \text{relu}(f\boldsymbol{X}_{i:i+l-1} + \boldsymbol{b}_{nf}) \quad (9)$$

where $\boldsymbol{b}_{nf}$ is a bias parameter, and relu is the rectified linear units following the specified transformation.

$$NF_l = \{NF_{l_1}, NF_{l_2}, \ldots, NF_{l_{n-l+1}}\} \quad (10)$$

After we calculate all the new features in turn, we can obtain a NF_l sequence over which we conduct a max pooling operation to get the maximum value of the new features $\hat{NF}_l = \max\{NF_l\}$. We also experimented with average pooling, but the results showed max pooling to be superior. Average pooling can capture more comprehensive features, but max pooling can capture the most important features, which is more useful for our task. We then use filters with different sentence window sizes l to get multiple features, and connect all $\hat{NF}_l$ to arrive at $\hat{NF}$. Using the CNNs, we extract n-gram features of the sentence, where N is the size of the sliding window. The words for judging uncertainty are given more weight in the attention mechanism, and max pooling helps us to focus on these words. Therefore, $\hat{NF}_l$, the output of CNNs, is an effective representation for uncertainty identification.

Finally, after a full connection neural network layer, we apply a softmax layer to produce the output,

$$p = \text{softmax}(\boldsymbol{W}_p \hat{NF} + \boldsymbol{b}_p) \quad (11)$$

We use cross-entropy against the correct labels as training loss,

$$L = -\sum_S \sum_C T_c(S) \log(p_c(S)) + \lambda l_2 \quad (12)$$

where C is the binary class of the sentence S, $T_c(S)$ is the binary value indicating whether the sentence S belongs to class C, while $p_c(S)$ is the prediction result of sentence S. l_2 is the L2 norm for regularization, and is the sum of the squares of all parameters, while λ is a parameter for extent to which l_2 should be calculated into the loss.

When we constructed our model, we tried to simulate the thought process of readers when identifying sentence uncertainty. Firstly, facing such a problem, people usually read through the sentence to understand each word and then return to the whole sentence; in our model, LSTM networks perform this task. To identify the level of uncertainty of the sentence, people then often pay attention to certain useful words, which is the purpose of our attention mechanism. Finally, people generally derive the result based on several prominent short sentences that carry certain or uncertain semantics. The treatment in our model is more comprehensive, as the CNNs' various sized sliding windows scan over the entire sentence. We believe the model is reasonable, and expect that it will achieve a performance equal to, or even beyond, a human reader in uncertainty identification.

4 Experiment

Experiments are conducted on both Chinese and English datasets. We first introduce our experimental setup, including descriptions of the datasets, alternative approaches for comparison, and some important preprocessing steps. Following this, we provide and discuss the experimental results.

4.1 Experiment setup

4.1.1 Dataset

To evaluate the performance of ANFU, we performed experiments on formal article texts and social media texts in both English and Chinese.

For English texts, we used the benchmark datasets of the CoNLL-2010 Shared Task 1[4], which targets the identification of sentences in texts expressing unreliable or uncertain information. Shared Task 1 consists of two datasets annotated by at least two linguists; one made up of biological articles and one made up of Wikipedia texts. We consider the Wikipedia dataset to represents social media texts. An overview of the experimental datasets is shown in Table 1. To better illustrate the

Table 1 Overview of CoNLL-2010 benchmark datasets.

Statistics	Number of Sentences	Number of average words in a sentence	Number of uncertain sentences	Ratio of uncertain sentences (%)	Number of uncertain cues	Number of average cues in a sentence
Biological article	19 544	23.30	3410	17.45	4423	1.30
Wikipedia texts	20 745	18.50	4718	22.74	6276	1.33

characteristics of the datasets, Table 1 also shows some statistics, including the number of uncertain sentences and the total amount of the occurrence of cue words in uncertain sentences. Of the two datasets, there are clearly more uncertain sentences and more uncertain cues in the Wikipedia texts. The two datasets are divided into training and test sets. There are 14 541 sentences for training and 5003 sentences for testing in the biological dataset, and 11 111 sentences for training and 9634 sentences for testing in the Wikipedia dataset.

In our experiment, we also evaluate our model on two Chinese datasets — a Chinese news dataset and a Chinese social media dataset. The Chinese news dataset is provided by Ji et al.[17] from data collected from Baidu News. To the best of our knowledge, this is the only Chinese news corpus for uncertainty identification, and consists of 10 000 sentences in total with 2858 of these classified as uncertain. In this dataset, a sentence is annotated as uncertain only when it contains a cue phrase.

Since there was no available Chinese social media corpus for uncertainty identification, we constructed the first Chinese social media dataset, which was collected from Sina Weibo during the Shanghai Expo. After data cleaning and extraction, we randomly selected 30 000 sentences to form our experiment dataset. We then manually annotated these sentences following CoNLL-2010 schema. A sentence is annotated as uncertain according to its semantics, regardless of the presence of cue phrases. To make the annotation labels credible, each sentence was judged by at least two people, producing a kappa value of 75.86%.

An overview of our experimental Chinese datasets is provided in Table 2, alongside some important statistics. There are two notable differences between the two datasets. First, the ratio of uncertain sentences in the social media dataset is larger than that in the news dataset, while the number of the average cues in the social media dataset are fewer than that in the news dataset. This justifies our assumption that cues tend to be absent in the informal style used in social media. Second, the average length of social media sentences is far longer than that of news sentences, indicating the former's greater complexity and difficulty to interpret.

For our experiment, we randomly chose 8000 sentences from the news corpora (with 2248 uncertain), and 24 000 sentences from the Weibo dataset (with 8798 uncertain) as the training set; the remainder were used as the test set.

4.1.2 Approaches for comparison

In our experiment, we chose several alternative approaches for comparison, including several important baselines and the state-of-the-art approach. Since our experimental datasets involve both English and Chinese, we selected different baselines accordingly.

Baseline 1: Reference [14] was the baseline used in the CoNLL task, and we also set it as Baseline 1 in our experiment.

Baseline 2: Reference [2] utilized the Hidden Markov Model (HMM) for uncertainty identification, and achieved the best results on the biological dataset in CoNLL-2010. We used it as Baseline 2 on the English datasets.

Baseline 3: Ji et al.[17] proposed a supervised method with lexical features as the first uncertainty detection method for Chinese texts, and we set it as Baseline 3.

Baseline 4: Reference [19] proposed an effective method based on cue-phrase features for Hungarian social media texts. Since our proposed approach also targets social media texts, we redesigned this method for processing Chinese texts and set is as Baseline 4.

CRK+ling: Tang[14] incorporated some external knowledge into a non-deep learning approach, which proved to be effective on the CoNLL-2010 dataset.

CNN Ex-Att: Reference [22] presented an attention architecture for uncertainty detection, using a CNN or

Table 2 Overview of the Chinese datasets.

Statistics	Number of Sentences	Number of average words in a sentence	Number of uncertain sentences	Ratio of uncertain sentences (%)	Number of uncertain cues	Number of average cues in a sentence
News dataset	10 000	30.72	2858	28.58	5084	1.79
Social media texts	30 000	41.14	11 071	36.90	11 618	1.05

RNN with an attention mechanism, which represents the state-of-the-art approach on the CoNLL2010 benchmark datasets. An external lexicon of seed cue words or phrases were also incorporated into the word embedding, and with the external knowledge their model performed well on English datasets.

ANFU: Our proposed attention-based LSTM-CNNs for uncertainty identification is evaluated with various configurations of the components in the neural networks, including CNN, RNN, RNN+ATT, and CNN+RNN.

For ease of comparison, we also adopted the official evaluation metrics of CoNLL-2010.

4.1.3 Preprocessing

Since our experimental datasets involve both English and Chinese corpora, we performed different preprocessing steps on each of them.

For Chinese, we used jieba (https://pypi.python.org/pypi/jieba/), an accurate and easy to use Python Chinese word segmentation module, for the word segmentation. We used gensim (http://radimrehurek.com/gensim/), a python package, to produce word vectors with deep learning via the word2vecs skip-gram model presented by Mikolov et al.[28] in 2013. We used 30 GB of Chinese texts, including Shanghai Expo Weibo and the Chinese news dataset of Ji et al.[17], to train a 100-dimension word vector. We also randomly selected 1000 sentences as a development set, on which the hyper-parameters of our model were tuned. Note that we do not remove the punctuation, as this can also carry meaning; for example, a "?" in a sentence usually indicates uncertainty.

For English, we use word vectors pre-trained by GloVe (https://nlp.stanford.edu/projects/glove/), which has 300-dimension vectors and a vocabulary of 2.2 million. The other parameters are set to the same as on the Chinese datasets.

During training, we used a two-layer RNN with attention mechanism — a deeper network was unnecessary because the two-layer network was efficient and performed well — and set the window size to 3, 4, 5, and 6 in our CNNs to extract the features. These window sizes were chosen because 3 to 6 consecutive Chinese words usually express a clear semantics. To avoid overfitting, we set the dropout parameter to 0.5.

4.2 Results and analysis

4.2.1 Results on English datasets

We firstly compared the effectiveness of uncertainty identification between our model and the alternatives on the CoNLL-2010 English benchmark datasets, with the results shown in Tables 3 and 4. From the experimental results, we found that our model ANFU outperformed the state-of-the-art on both datasets, with F1-measures of 70.02% and 87.21%.

When comparing the results shown in Tables 3 and 4, we found that almost all models performed better on the biological dataset than on the Wikipedia dataset except for Baseline 1 on all the metrics. We also found that CNN Ex-Att and ANFU outperformed the other models on the Wikipedia dataset. This is due to the diversity of expressions in Wikipedia texts, with certain words, such as '*high*', '*groups*', and '*great*', that are regarded as certain in formal texts, sometimes carrying the semantics of uncertainty in Wikipedia texts. In this case, the model of deep learning can obviously achieve better results than those approaches based on cue-phrases.

In addition, on the Wikipedia dataset, the size of the training set is almost equal to that of the testing set, making it difficult for some learning algorithms to capture all the features for uncertainty identification.

4.2.2 Results on Chinese datasets

Tables 5 and 6 show the experimental results on the

Table 3 Results on the English Wikipedia dataset.

Model	Precision	Recall	F1-measure
Baseline 1	0.7203	0.5429	0.6191
Baseline 2	0.6691	0.6128	0.6397
CRK+ling	**0.8228**	0.4136	0.5505
CNN Ex-Att	-	-	0.6752
ANFU	0.7776	**0.6368**	**0.7002**

Table 4 Results on the English biological dataset.

Model	Precision	Recall	F1-measure
Baseline 1	0.6907	**0.9101**	0.7854
Baseline 2	0.7332	0.8835	0.8015
CRK+ling	0.8712	0.8646	0.8679
CNN Ex-Att	-	-	0.8557
ANFU	**0.8748**	0.8695	**0.8721**

Table 5 Results on the Chinese social media dataset.

Model	Precision	Recall	F1-measure
Baseline 3	0.6754	0.6502	0.6625
Baseline 4	0.5271	0.5425	0.5347
CNN	0.7081	0.6286	0.6659
RNN	0.7127	0.6800	0.6959
RNN+ATT	0.7681	0.7186	0.7425
CNN+RNN	0.7662	0.7029	0.7331
ANFU	**0.7784**	**0.7856**	**0.7819**

Table 6 Results on the Chinese news dataset.

Model	Precision	Recall	F1-measure
Baseline 3	0.7024	0.7082	0.7053
Baseline 4	0.6235	0.6620	0.6422
CNN	0.5928	0.6230	0.6075
RNN	0.6343	0.6808	0.6567
RNN+ATT	0.7090	0.7289	0.7188
CNN+RNN	0.7010	0.7157	0.7083
ANFU	**0.7414**	**0.7377**	**0.7395**

Chinese news dataset and Chinese social media dataset, respectively.

From the results shown in Tables 5 and 6, we found that ANFU (CNN+RNN+ATT) outperformed both baselines, and yielded F1-measure scores of 78.19% and 73.95% on the social media dataset and news dataset, respectively, marking a 41.37% and 4.8% improvement over Baseline 3. These results show that ANFU can perform well on both news data and social media data, while providing a much larger improvement on social media data.

Note that among the 2858 uncertain sentences in the news dataset, there are only 23 uncertain sentences (0.8%) not containing cue phrases. This means that ANFU does not give a large improvement over the baselines. In the social media dataset, however, 22.28% of uncertain sentences do not contain cue phrases, leading both baselines to fail to identify the uncertain sentences. Because ANFU does not rely on cue phrases, the uncertain semantics can still be captured by the attention mechanism. There were 1.05 cue-phrases in the social media dataset and 1.79 cue-phrases in news dataset as shown in Table 2. In Tables 5 and 6, it was observed that ANFU achieved 18.82% improvement on Weibo dataset, but only 4.8% improvement on news dataset. This can be explained by ANFU's better understanding of uncertain semantics in a substandard shape or even omitted from social media texts.

To summarize, the experimental results in Tables 3 and 5 show that ANFU outperformed the state-of-the-art in terms of F1-measure on both the Wikipedia dataset and Chinese social media dataset. Furthermore, the results showed that the attention mechanism was helpful, since both RNN+ATT and ANFU outperformed their counterparts without an attention mechanism. In analyzing the results of different models, we considered that RNNs could capture more global features, so had a high recall by understanding the general idea of a sentence. CNNs and the attention mechanism could grasp the most important features for uncertainty identification to obtain a high precision, while the attention mechanism was more effective. As a result, the combination of these achieved the best performance.

To illustrate the effectiveness of our attention mechanism, sentence "Under the similar odds, away winning probability of the home team is very small." taken from the social media dataset is visualized with attention weights in Fig. 2, where a darker color means a higher weight, indicating uncertainty. In this instance, there is no cue-phrase in the sentence, so it is unlikely to be identified by either baseline method. Using ANFU, however, the word "odds" with its implicit uncertain semantics can be captured by attention mechanisms, even though it is not a cue-phrase, and hence the sentence is determined as uncertain.

We also compare the performance of accuracy with different training steps on different datasets, as shown in Fig. 3. By using ANFU, the accuracy will be achieved more than 0.8 on both news and social media datasets, but on the latter one it would reach a higher accuracy with less steps, which prove that the social media dataset will be benefitted from our proposed method.

Finally, we examined the effect of sentence length on the accuracy of our model, because the length of sentences often has a great influence on text

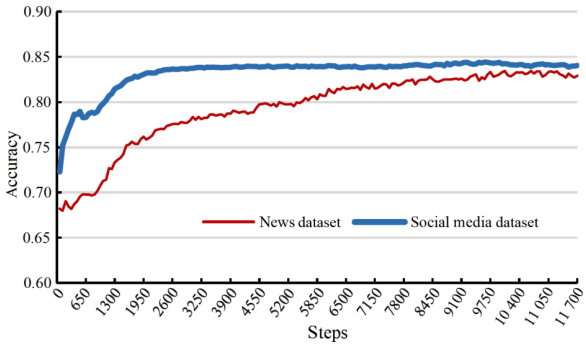


Fig. 3 Accuracy of different training steps.

Chinese sentence	相近 (similar)	赔率 (odds)	下 (under)	西甲 (Liga)	主队 (home team)	客胜 (away win)	概率 (prob.)	很小 (very small)	。 (.)
English translation	Under the similar odds, away winning probability of the home team is very small.								

Fig. 2 Visualization of uncertainty attention over words.

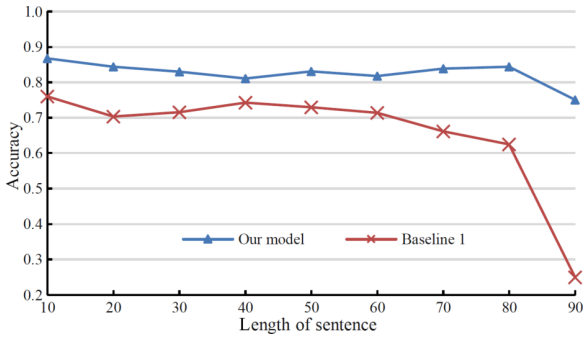


Fig. 4 Accuracy for different sentence lengths.

classification tasks. The length of a sentence here is calculated by the number of words, not the number of characters. We divided the sentences into 9 groups at intervals of 10 words. The accuracy values were drawn at the end of each group, such as the position at 10, 20, etc. Figure 4 shows a comparison of the results between Baseline 1 and our model. We found that the accuracy of our model did not decrease with the increase of sentence length, whereas the accuracy of Baseline 1 dropped with the increase of the sentence length, and dropped sharply when the sentence length reached 80. Since there is a high likelihood of long sentences in social media, our model's ability to cope with long sentences makes it trustworthy for social media text.

5 Conclusion and Future Work

This paper proposes an attention-based neural framework for uncertainty identification on Chinese social media texts, named ANFU. ANFU uses attention-based LSTM networks to focus on the words, regardless of the presence of cue-phrases, that have a decisive effect on the uncertain semantics of the sentence, without resorting to extra knowledge or external NLP components. The convolutional neural networks of ANFU capture the most important semantic information for uncertainty identification.

Experiments were conducted on four datasets, made up of 2 English benchmark datasets from the CoNLL-2010 task of uncertainty identification, and 2 Chinese datasets from Weibo (Chinese Microblogging platform) and Chinese news. Experimental results showed that our proposed ANFU approach outperformed the-state-of-the-art on all of these datasets in terms of F1-measure. More importantly, 41.37% and 13.10% improvements were achieved over the baselines on English and Chinese social media datasets, respectively, showing the particular effectiveness of ANFU on social media texts.

In the future, we will expand the social media dataset and look to make further classification of the uncertain sentences into different types of uncertainty.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (Nos. 61502115, 61602326, U1636103, U1536207, and 61672361), the Fundamental Research Fund for the Central Universities (No. 3262019T29), and the Joint Funding for Capital Universities (No. SKX182010023).

References

[1] G. Szarvas, V. Vincze, R. Farkas, G. Mora, and I. Gurevych, Cross-genre and cross-domain detection of semantic uncertainty, *Computational Linguistics*, vol. 38, no. 2, pp. 335–367, 2012

[2] X. J. Li, W. Gao, and J. W. Shavlik, Detecting semantic uncertainty by learning hedge cues in sentences using an HMM, in *Proceeding of the Special Interest Group on Information Retrieval (SIGIR'14)*, Gold Coast, AUS, 2014, pp. 89–107.

[3] R. Farkas, V. Vincze, G. Mora, J. Csirik, and G. Szarvas, 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 (CoNLL'10)*, Uppsala, Sweden, 2010, pp. 1–12.

[4] Y. T. Wei, X. G. You, and H. Li, Multiscale patch-based contrast measure for small infrared target detection, *Pattern Recognition*, vol. 58, no. 1, pp. 216–226, 2016.

[5] Z. Y. Wei, J. W. Chen, W. Gao, B. Y. Li, and L. J. Zhou, An empirical study on uncertainty identification in social media context, in *Proceedings of the Association for Computational Linguistics (ACL'13)*, Minneapolis, MN, USA, 2013, pp. 58–62.

[6] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint, arXiv: 1409.0473, 2014.

[7] Y. Kim, Convolutional neural networks for sentence classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (EMNLP'14)*, Doha, State of Qatar, 2014, pp. 1746–1751.

[8] R. Morante and C. Sporleder, Modality and negation: An introduction to the special issue, *Computational Linguistics*, vol. 38, no. 2, pp. 223–260, 2012.

[9] M. Light, X. Y. Qiu, and P. Srinivasan, The language of bioscience: Facts, speculations, and statements in between, in *Proceedings of the BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, Boston, MA, USA, 2004, pp. 17–24.

[10] W. W. Chapman, D. Chu, and J. N. Dowling, Context: An algorithm for identifying contextual features from clinical text, in *Proceedings of the ACL Workshop on BioNLP*, Prague, Czechoslovakia, 2007, pp. 81–88.

[11] B. Medlock, Exploring hedge identification in biomedical

literature, *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 636–654, 2008.

[12] E. Fernandes, C. Crestana, and R. Milidiu, Hedge detection using the RelHunter approach, in *Proceedings of the 14th Conference on Computational Natural Language Learning-Shared Task (CoNLL'10)*, Uppsala, Sweden, 2010, pp. 64–69.

[13] X. X. Li, J. P. Shen, X. Gao, and X. Wang, Exploiting rich features for detecting hedges and their scope, in *Proceedings of the 14th Conference on Computational Natural Language Learning-Shared Task (CoNLL'10)*, Uppsala, Sweden, 2010, pp. 78–83.

[14] B. Z. Tang, X. L. Wang, X. Wang, B. Yuan, and S. X. Fan, 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, (CoNLL'10)*, Uppsala, Sweden, 2010, pp. 13–17.

[15] S. D. Zhang, H. Zhao, G. D. Zhou, and B. L. Lu, Hedge detection and scope finding by sequence labeling with normalized feature selection, in *Proceedings of the 14th Conference on Computational Natural Language Learning-Shared Task, (CoNLL'10)*, Uppsala, Sweden, 2010, pp. 92–99.

[16] E. Velldal, Detecting uncertainty in bio-medical literature: A simple disambiguation approach using sparse random indexing, in *Proceedings of the International Symposium on Semantic Mining in Biomedicine (SMBM'10)*, Cambridge, UK, 2010, pp. 75–83.

[17] F. Ji, X. P. Qiu, and X. J. Huang, A research on Chinese uncertainty sentence recognition, in *Proceedings of the 16th China Conference on Information Retrieval (CCIR'10)*, Harbin, China, 2010, pp. 595–601.

[18] V. Vincze, Uncertainty detection in natural language texts, PhD dissertation, Stanford University, Palo Alto, CA, USA, 2015.

[19] V. Vincze, Detecting uncertainty cues in Hungarian social media texts, in *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, Osaka, Japan, 2016, pp. 11–21.

[20] Y. Zhang and B. C. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, arXiv preprint, arXiv: 1510.03820, 2015.

[21] Z. C. Yang, D. Y. Yang, C. Dyer, X. D. He, A. Smola, and E. Hovy, Hierarchical attention networks for document classification, in *Proceeding of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'16)*, San Diego, CA, USA, 2016, pp. 1480–1489.

[22] H. Adel and H. Schutze, Exploring different dimensions of attention for uncertainty detection, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (ECACL'17)*, Valencia, Spain, 2017, pp. 22–34.

[23] S. Hochreiter and J. Schmidhuber, Flat minima, *Neural Computation*, vol. 9, no. 1, pp. 1–42, 1997.

[24] M. Sundermeyer, R. Schlüter, and H. Ney, LSTM neural networks for language modeling, in *Proceedings of the Interspeech*, Portland, OR, USA, 2012, pp. 194–197.

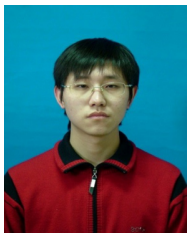
[25] K. S. Yao, T. Cohn, K. Vylomova, K. Duh, and C. Dyer, Depth-gated recurrent neural networks, arXiv preprint, arXiv: 1508.03790, 2015.

[26] A. Graves, Generating sequences with re-current neural networks, arXiv preprint, arXiv: 1308.0850, 2013.

[27] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, Natural language processing (almost) from scratch, *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2493–2537, 2011.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representa-tions in vector space, arXiv preprint, arXiv: 1301.3781, 2013.

Xu Han received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2011. She is currently an assistant professor of the Information Engineering College, Capital Normal University. Her areas of research include natural language processing, social computing, and cloud computing. She has served as an reviewer of several international journals, and as a member of Chinese Information Processing Society (CIPS).



Binyang Li received the PhD degree from the Chinese University of Hong Kong in 2012. He is currently an associated professor and a master supervisor of the School of Information Science and Technology, University of International Relations. His areas of research include natural language processing and social computing. He has published over 50 refereed papers. He has served as an reviewer of several international journals, and served as a committee member of several conferences.

Zhuoran Wang received the PhD at the University College London in 2009. He is currently the founder and CEO of Tricorn (Beijing) Technology Co., Ltd. His research interests include spoken dialogue systems, natural language processing, and machine learning. He has published over 30 refereed papers.