Journal of Biomedical Informatics 60 (2016) 334–341

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models

Shaodian Zhang [a], Tian Kang [a], Xingting Zhang [b], Dong Wen [b], Noémie Elhadad [a], Jianbo Lei [b,*]

[a] *Department of Biomedical Informatics, Columbia University, New York, USA*
[b] *Center for Medical Informatics, Peking University, Beijing, China*

ARTICLE INFO

Article history:
Received 11 November 2015
Revised 9 February 2016
Accepted 12 February 2016
Available online 26 February 2016

Keywords:
Speculation detection
Natural language processing
Chinese NLP
Clinical NLP
Word embedding
Word segmentation

ABSTRACT

Speculations represent uncertainty toward certain facts. In clinical texts, identifying speculations is a critical step of natural language processing (NLP). While it is a nontrivial task in many languages, detecting speculations in Chinese clinical notes can be particularly challenging because word segmentation may be necessary as an upstream operation. The objective of this paper is to construct a state-of-the-art speculation detection system for Chinese clinical notes and to investigate whether embedding features and word segmentations are worth exploiting toward this overall task. We propose a sequence labeling based system for speculation detection, which relies on features from bag of characters, bag of words, character embedding, and word embedding. We experiment on a novel dataset of 36,828 clinical notes with 5103 gold-standard speculation annotations on 2000 notes, and compare the systems in which word embeddings are calculated based on word segmentations given by general and by domain specific segmenters respectively. Our systems are able to reach performance as high as 92.2% measured by *F* score. We demonstrate that word segmentation is critical to produce high quality word embedding to facilitate downstream information extraction applications, and suggest that a domain dependent word segmenter can be vital to such a clinical NLP task in Chinese language.

1. Introduction

An increasing amount of computerized clinical data is becoming available with the adoption of electronic health records (EHRs). Natural language processing (NLP) and information extraction techniques have been critical parts of the pipeline to automate the EHR data structuring, data mining and knowledge discovery, and have become an active research field in biomedical informatics. For example, researchers have made significant progress on named entity recognition (NER) from clinical and biomedical texts, whose aim is to detect the boundaries, to identify the categories of clinical entities, and to map them to concepts in standardized terminologies [1–3].

One of the issues that need to be addressed along with NER in clinical information extraction is to detect cues of speculations (sometimes also referred to as hedges, or uncertainties), which has vital impacts on the credibility of statements or hypotheses in contents [4]. Linguistic speculations are used when uncertainty is expressed, such as in the sentence “the patient may have a UTI”, where “may” is the cue responsible for the uncertainty. Identification of such cues is critical to downstream applications like knowledge discovery, question answering, and predictive modeling of diseases.

Most existing methods for speculation detection are developed for English text. In recent years, hospitals in China have been rapidly deploying EHR systems, which generate a great amount of clinical data. Efforts have been made in the research community to construct NLP components for Chinese clinical notes [5–8], but to our best knowledge there is no established system identifying clinical speculations.

Secondary use of EHR data in China requires NLP pipelines tailored to clinical language in Chinese, which is dramatically different from clinical notes in English and makes migration of existing NLP systems impractical. One of the primary difference between Chinese and English NLP is that most Chinese NLP systems need to begin with word segmentation, an unnecessary step in its English counterpart [9]. The need for word segmentation, which also exists in other languages like Arabic [10], Hebrew [11], and Japanese [12], creates additional challenges for

* Corresponding author at: Center for Medical Informatics, Peking University, No. 38 Xueyuan Rd, Haidian District, Beijing 100191, China. Tel.: +86 (10) 8280 5901; fax: +86 (10) 8280 5900.
E-mail address: jblei@hsc.pku.edu.cn (J. Lei).

http://dx.doi.org/10.1016/j.jbi.2016.02.011
1532-0464/© 2016 The Authors. Published by Elsevier Inc.

automated NLP pipelines including named-entity recognition and other components (e.g., parsing, information extraction, etc.). Linguistic resources and tools have been created for Chinese word segmentation [9,13,14], but they were built for general purposes and were not tailored to handling clinical texts. As such, one question worthy of investigation is the impact of word segmentation on our task at hand, speculation detection in Chinese clinical texts.

In recent years, a novel type of feature has been proposed for NLP tasks such as text classification, parsing, and sentiment, namely, word embeddings [15,16]. Word embeddings are usually learned via neural networks. Instead of using each word in an NLP task as a feature, words are represented as vectors which could encode rich contextual information. In a broader scope, word representation based on distributional semantics has been exploited in a wide range of clinical NLP tasks such as named entity recognition [17–19] and lexicon expansion [20]. An important motivation of using word embedding in many information extraction tasks, including speculation detection, is that it can learn semantic representations of words from a large unannotated corpus, while the training corpus for the task itself is usually much smaller and more sparse. This can be particularly helpful in clinical NLP because large amount of clinical notes can be available but manual annotations by medical experts are usually too costly. As such, the use of word embeddings as a feature in a speculation detection algorithm is another important question to investigate in our study.

In Chinese texts, word segmentation must be carried out before training a word embedding model. Unfortunately, high-quality segmentation is not always available when the text is highly domain specific, such as for clinical notes. As a substitute for word embeddings, character embeddings have been exploited in applications [21,22], which build feature vectors for each character instead of word. However, such substitution may not be ideal, since words, instead of characters, are units of language that carry meanings upon which semantic representations should be built. For Chinese clinical NLP, the two questions of interest, impact of word segmentation algorithms and impact of embedding features, are inter-related.

In this paper, we propose a sequence labeling based system identifying speculations from Chinese clinical text. To investigate our two questions, we experiment with four types of features: bag of characters, character embedding, bag of words, and word embedding. For the latter two that rely on word segmentation, we carry out experiments with a general-purpose Chinese word segmenter, and a segmenter specially trained on clinical notes, respectively. Our system, to our best knowledge, is the first speculation detection system for Chinese clinical notes. We compare the effectiveness of these four groups of features, and demonstrate that a domain-dependent word segmenter and embedding features can be helpful in such a clinical information extraction task.

1.1. Related work

In the biomedical domain, speculation detection has been an organic component in clinical NLP as early as first wave of automated text processing systems such as in [4]. These early systems usually rely on hand-crafted rules or grammatical patterns to identify statements with uncertainty in clinical notes. Since 2000, speculation detection from biomedical texts, especially scientific literature, has been flourishing thanks to the emergence of shared linguistic corpora and pioneer works [23,24]. The BioScope corpus, in particular, provides a benchmark for speculation detection as well as negation detection in several systems [25–27]. The corpus comprises both biomedical literature texts and clinical notes. Part of the corpus was also adopted for shared tasks on hedge detection, such as CoNLL-2010 which also prompted an enthusiastic response from the international research community [27]. Utilizing manually labeled corpora, machine learning-based speculation cue detectors have been developed, which leverage SVM [25,28,29], CRF sequence labeling [30–32], or decision trees [25]. Semi-supervised learning [33] or hybrid methods [34,35] were also developed to identify the hedges. It is noteworthy that many of the works detecting speculative cues also identify linguistic scopes of these cues, given that BioScope and CoNLL-2010 both include scope annotation in addition to cues. Scope finding, compared to cue detection, is a more challenging task which usually requires deeper syntactic analysis over the text [27]. While most of the previous works focus on biomedical speculations, other research has focused on identifying hedges from general scientific literature [36] or Wikipedia text [27,37].

In Chinese NLP, several negation and speculation detection systems have recently been developed, but primarily on scientific literature [38–40] and general news texts [41]. In the clinical domain, various NLP systems for Chinese clinical text have been created, such as those for word segmentation [8], named entity recognition [6], information extraction [7], and term alignment [42]. No speculation detection system is developed specifically for Chinese clinical notes so far, to our best knowledge.

2. Material and methods

2.1. Dataset and annotations

One month of admission notes and discharge summaries were collected from the EHR database of Peking Union Medical College Hospital, resulting in 36,828 notes in total from 17 departments and clinics. After excluding incomplete notes, 2000 clinical notes (1000 admission notes and 1000 discharge summaries) were randomly sampled for this study. The 2000 notes were de-identified manually. Two medical doctors (DW and XZ) and one clinical informaticist (SZ) drafted the first version of the annotation guideline, which was strongly inspired by the "minimal unit" principle in the guideline of the BioScope corpus [24]. Then a pilot annotation on 10 discharge summaries, corresponding to approximately 30 speculation cues, was carried out by the two medical doctors to refine the annotation guideline. It is noteworthy that only speculative cues (keywords) were annotated, while linguistic scopes of these cues were not taken into consideration for this study. Our definition of speculation cues also includes keywords and phrases expressing vagueness, such as the "more than" in "more than 10 year history of diabetes", since they occur very frequently and brings uncertainties to the facts as well. Furthermore, we asked the annotators to pay special attention to hedge cues for four categories of information: disease and syndrome, symptom and sign, treatment and drug, and laboratory test, based on the named entity recognition results which was described in our previous work on the same data set [6]. Following are several example sentences with speculations annotated (underline and italic):

肺部感染 可能性大 *Most likely* lung infection
疑似溃疡性结肠炎 *suspicious* of ulcerative colitis
肾功能大致正常 Renal function *roughly* in the normal range

The two medical doctors (DW and XZ) then double annotated 80 notes to calculate the inter-rater agreement. An agreement of 0.76 measured by Kappa [43] was reached. The two annotators then resolve all disagreements and refine the guideline. Our final annotation guideline is given in Appendix A. The remaining 1920 notes were evenly split and coded by single annotator only. In total, 5103 speculation cues were identified, in which 2558 modifies diseases and diagnosis, 1134 modifies symptoms, 960 modifies laboratory tests, and 451 modifies treatment and drug.

2.2. Baseline

The annotators were asked to keep track of and generalize annotation rules heuristically during the annotation process. The common part of the rule sets given by the two annotators, which consists of a list of 31 speculation cues with 16 rules of constraints (Table 1), was implemented with regular expressions and used as the baseline system in our experiments. Such rules include ones like "when "余 (more than)" appears right after a number, like it does in "10 余 (more than 10)", it should be annotated as a speculation cue".

2.3. CRF-based sequence labeling

The task of identifying speculative cues can be cast as a sequence labeling problem. Conditional Random Fields (CRF), an established method for numerous information extraction tasks including in clinical domain, was adopted [1,44–46]. In our task, we used the classical 'BIO' notations to represent the boundaries of cues. The following example shows how a sentence with speculation is labeled, in which *B-spec* represents the beginning character of a speculation cue, *I-spec* represents being inside a cue, and O represents being outside of speculation cues:

> 肺/O 部/O 感/O 染/O 可/*B-spec* 能/*I-spec* 性/*I-spec* 大/I-spec *Most likely* lung infection
>
> 糖/O 尿/O 病/O 史/O10/O 余/*B-spec* 年/O *More than* 10 year history of diabetes

We only used "B-spec", "I-spec", and "O" as labels and did not distinguish cues by entity types they modify. We rely on the open source CRF++ tool [47] for the training and prediction with default parameters for our experiments.

2.4. Features

To investigate our research questions (impact of word segmentation quality and use of embeddings as features for speculation detection), we fed different types of features into a series of CRF labelers. For each labeler, the input is a sentence and the output is a BIO-type sequence.

Table 1
Cues and constraints selected by the two annotators. Cues and constraints listed in this table are used as the baseline system to match speculation cues from the original text.

Cues without constraints	Cues with constraints
不除外 (cannot exclude), 不排除 (cannot exclude), 待排 (to be excluded), 待排除 (to be excluded), 待除外 (to be excluded), 可能大 (very likely), 可能性大 (very likely), 推测 (speculate), 也许 (probably), 疑似 (suspicious), 不能完全除外 (cannot exclude), 可疑 (suspicious), 稍 (a little), 略 (a little), 时有 (occasionally)	1. ? after disease names; 2. 拟 (probably) before 诊断 (diagnose); 3. 约 (around) before numbers; 4. 左右 (around) after numbers; 5. 余 (more than) after numbers; 6. 许 (more than) after numbers; 7. 考虑 (consider) when appearing with 可能 (possibility); 8. 考虑 (consider) when appearing with 可能性 (possibility); 9. 可能 (possible) after disease names; 10. 大约 (around) before numbers except time; 11. 大概 (around) before numbers except time; 12. 基本 (overall) before 正常 (normal); 13. 大致 (roughly) before 正常 (normal); 14. 偶发 (occasionally happen) before or after symptoms; 15. 偶有 (occasionally have) before or after symptoms; 16. 偶见 (occasionally see) before or after symptoms

The first system in the study relies on the original input of individual characters as features. No word segmentation were carried out for this system. Thus, suppose that the vocabulary in the training set has V_c characters, and each unique character v_i has a unique index i. For each character to be labeled, a one-hot feature vector f_i of V_c dimensions will be created: $f_i = \langle 0, \ldots, 1, \ldots, 0 \rangle$, where value of the ith element is 1 and all other elements are zero.

The second type of feature we use is bag of words. State-of-the-art Chinese word segmenters can achieve around 95% accuracy on general text, which is sufficient for most applications [9]. However, in the clinical domain, word segmenters for general purposes usually fail as other NLP systems do, since content in clinical notes is highly domain specific and dependent, no matter what the language is [48]. As such, a segmentation tool needs to be re-trained on the clinical notes in order to be effective in providing accurate results for downstream word embedding model. In this study, we try separately two segmenters to generate bag of words representations, one taken directly from a state-of-the-art general purpose Chinese NLP pipeline, and the other specifically trained on clinical notes. Section 2.6 will describe how we trained the domain-specific word segmenter.

In order to enrich the feature set, we expand the space by adding distributional representations of characters and/or words, leading to the third and fourth types of features we use: character embedding and word embedding. More precisely, we use vectorized embeddings to get richer representations of linguistic units (characters or words), which is basically a parameterized function mapping units to high-dimensional vectors [15]. Intuitively, embedding models encode hidden linguistic information that a word/character can convey in different context into a vector of certain dimensions. Previous research shows that the embedding space is much more powerful than the one-hot representation (e.g., bag-of-words), and can make breakthroughs in many NLP tasks as it conveys more semantic meanings and is particularly useful in overcoming sparsity [15,16]. In our task, for each word or character t, depending on whether the embedding is word level or character level, a vector with fixed number of dimensions (N) will be calculated like the following: $w(t) = \langle 0.2, -0.4, 0.7, \ldots \rangle$, based on unsupervised parameter estimation on the unlabeled corpus. The vector can then be used directly as an N-dimensional feature vector for the CRF labeler. Similarly with bag of words, word embedding can be calculated based on different segmentation results, given by a general-purpose segmenter or by a domain specific one.

Fig. 1 illustrates how the four types of feature vectors described above – bag of character, bag of words, character embedding, and word embedding – are calculated for a snippet of text. For bag of words and bag of characters, one-hot representations are calculated for units of interest, i.e., characters or words. Character embedding and word embedding are obtained by training neural network models on the corpus, relying on the original sequence of characters and the segmented sequence of words, respectively. In the case of bag of characters and bag of words, the individual dimensions correspond to the characters and words in the corpus, and thus are quite large. In the embedding representations, the number of dimensions is much lower, but as a trade-off, each individual dimension is harder to interpret for humans. We also note that the weights under the embedding representations are learned unlike in the bag of word and character representations, where each weight simply indicates presence or absence in the input.

2.5. The workflow

For each type of feature, two systems have been implemented: one only using the vector of current unit as features (unigram), the other including vectors from both current unit and previous unit

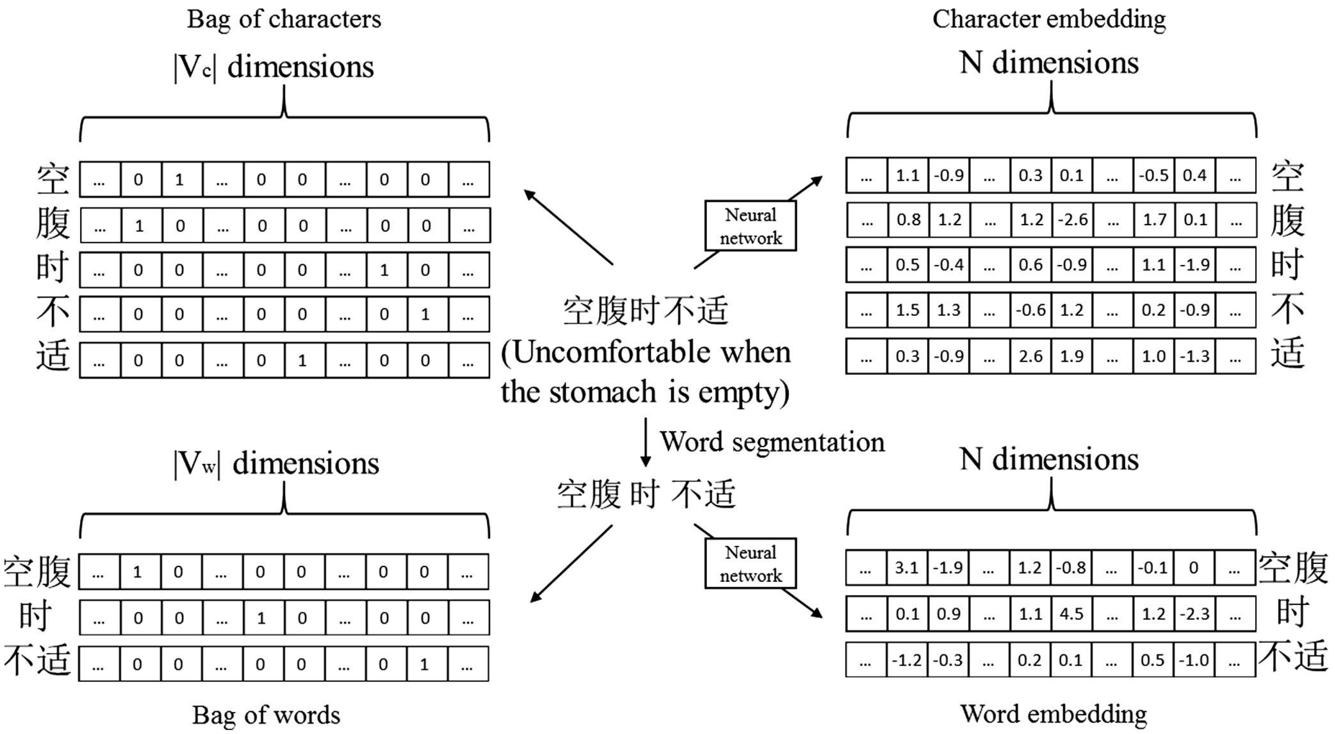

Fig. 1. Different representations for the feature vectors for the four types (bag of character, bag of words, character embedding, and word embedding) for a given text snippet. Upper-left are one-hot vectors in bag-of-character representations, and $|V_c|$ is the total number of possible characters. Lower-left are one-hot vectors in bag-of-words representation, and $|V_w|$ is the total number of words after word segmentation. Upper-right and lower-right are character embedding and word embedding representations, after trained by two independent neural networks. N is a hyper-parameter and is set as 100 in this study.

(bigram). The entire unlabeled dataset extracted (36,828 notes) were used to learn the word/character embeddings. We used the word2vec tool's Continuous Bag of Words (CBOW) model to train all embeddings [49], and set vector size $N = 100$, iteration number 20 and all other parameters default.

An overall workflow for all systems is given in Fig. 2. Twelve systems based on four types of features are implemented, in addition to the rule-based baseline. The BOC systems use one hot encodings of characters as features. C2V-unigram and C2V-bigram system use character embedding. BOW-G and W2V-G systems are bag of words and word embedding systems based on a word segmentation given by the Stanford word segmenter [50], which is trained on Chinese news text, and BOW-Ds and W2V-Ds are the systems using word embedding based on word segmentation given by a CRF segmenter trained on the same type of clinical notes.

2.6. Supervised word segmentation for Chinese clinical notes

A domain specific word segmenter is needed to generate features for BOW-Ds and W2V-Ds systems described above. To train such a segmenter, 100 notes were randomly sampled and annotated manually by the same annotators as for the speculation detection task. An agreement of 0.74 has been achieved before the annotators resolve disagreements. A CRF segmenter is then trained and evaluated on the annotated data. Finally, the segmenter is applied to the entire 2000 notes with speculation annotation and the results are taken by to BOW-Ds and W2V-Ds systems for feature extraction.

3. Results

For each system, we carry out a 5-fold cross validation and report average performance in precision, recall, and F-score. Detailed performance for all systems is given in Table 2. For each system, we re-sampled the 5 folds of data for 5 times, resulting in 25 folds in total with 25 scores. We calculated the 95% confidence intervals by using these 25 performance scores, assuming they are normally distributed. The confidence intervals can be used to measure whether differences among systems are indeed significant.

Overall, all the CRF-based systems outperform the rule based baseline significantly, and all bigram systems reach F scores of over 90, except BOC-bigram and BOW-G-bigram. Several observations can be made. First, for all types of feature, adding bigram helps to increase the performance significantly. Second, most word-based systems outperform their character-based counterparts, only if domain-specific word segmentation is used. The general-purpose segmenter may undermine the effectiveness of using words, making BOW-Gs and W2V-Gs perform poorer than BOCs and C2Vs. Third, embedding representation can enhance the system performance, compared with one-hot bag of characters or bag of words. The differences are particularly significant with word-based systems (i.e., W2Vs vs. BOWs). Finally, the best system W2V-D-bigram, which takes all advantages of bigram, embedding, and a domain specific word segmentation, outperform all other ones with significant differences.

The type of segmenter used made a difference in the overall task of speculation detection. Before discussing their impact, we report the evaluation of the segmenters themselves on our dataset. The Stanford word segmenter was used in W2V-Gs and BOW-Gs and the CRF word segmenter was used in W2V-Ds and BOW-Ds. We cross-validated the CRF word segmenter using the 100 annotated admission notes with 5 folds, and for each fold we also evaluated the performance of Stanford segmenter. For both tools, we use *Contemporary Chinese Dictionary (CCD)* [51] as the dictionary to evaluate IV (in vocabulary) and OOV (out of vocabulary) performance. The rationale behind is that CCD lists common Chinese words and excludes domain specific clinical terms. As such, IV could be regarded as how good the word segmenters identify

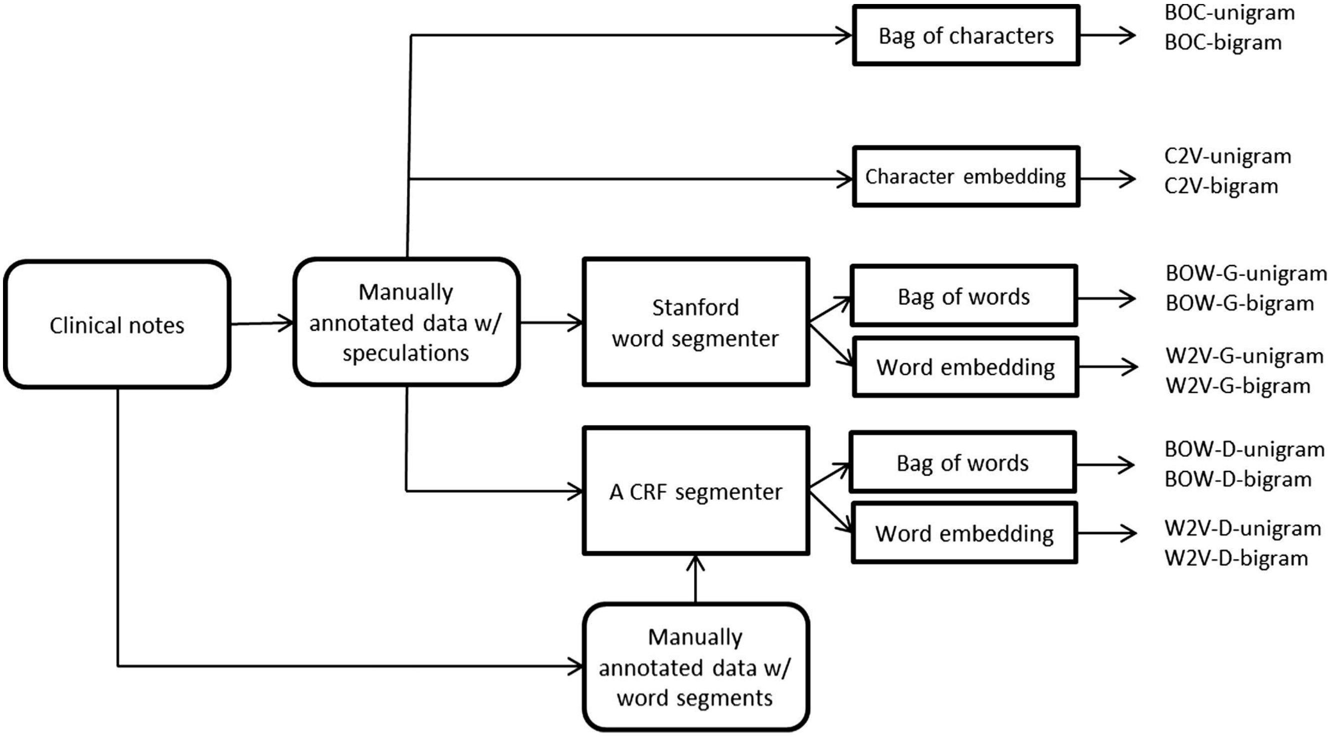

Fig. 2. workflow to obtain the twelve systems. BOCs are the systems using bag of characters as features. C2V systems use character embedding. BOW-Gs and W2V-Gs use bag of words and word embedding based on a word segmentation given by the Stanford word segmenter. BOW-Ds and W2V-Ds are the systems using bag of words and word embedding based on word segmentation given by a CRF segmenter trained on clinical notes.

Table 2
Performance of all systems measured by precision, recall, and *F* score. 95% confidence intervals are included in brackets. Descriptions of systems refer to Fig. 1. Best precision, recall and *F* are in bold.

	Precision (95% CI)	Recall (95% CI)	*F* (95% CI)
Baseline (rule based)	59.1 (±0.7)	84.5 (±0.6)	69.5 (±0.6)
BOC-unigram	85.1 (±0.5)	85.6 (±0.4)	85.4 (±0.5)
BOC-bigram	91.5 (±0.6)	88.1 (±0.6)	89.8 (±0.6)
C2V-unigram	86.7 (±0.5)	85.5 (±0.5)	86.1 (±0.6)
C2V-bigram	92.4 (±0.5)	**90.2 (±0.4)**	91.3 (±0.4)
BOW-G-unigram	83.7 (±0.6)	82.3 (±0.5)	82.9 (±0.6)
BOW-G-bigram	87.4 (±0.5)	86.9 (±0.4)	87.1 (±0.4)
BOW-D-unigram	84.9 (±0.4)	86.0 (±0.5)	85.4 (±0.5)
BOW-D-bigram	91.2 (±0.3)	89.5 (±0.3)	90.3 (±0.3)
W2V-G-unigram	85.6 (±0.8)	85.1 (±0.6)	85.4 (±0.7)
W2V-G-bigram	91.8 (±0.5)	89.9 (±0.4)	90.9 (±0.5)
W2V-D-unigram	89.9 (±0.7)	84.5 (±0.6)	87.1 (±0.6)
W2V-D-bigram	**94.5 (±0.5)**	90.1 (±0.3)	**92.2 (±0.4)**

Table 3
Overall, In-Vocabulary, and Out-of-Vocabulary performance of Stanford segmenter and our CRF segmenter trained on admission notes, measured by *F* score. Vocabulary used is the *Contemporary Chinese Dictionary (CCD)*. IV can be seen as how good the word segmenters identify common word and OOV approximates how good the segmenters handle clinical terms.

	Overall	IV	OOV
Stanford	69.0	84.9	43.8
Ours	83.1	87.6	74.1

common word, and OOV approximates how good the segmenters handle clinical terms. Table 3 gives overall, IV, and OOV *F* scores for the two systems.

The results show that Stanford word segmenter trained on general Chinese news text cannot handle clinical notes well, especially on clinical terms, while training our CRF segmenter on the annotated admission notes of the same genre can boost the performance of word segmentation. The results are not surprising, but help confirm that the performance increase from BOW-Gs and W2V-Gs to BOW-Ds and W2V-Ds we showed previously is indeed the effect of a better word segmenter, especially a better segmentation of clinical terms.

4. Discussions

4.1. Findings and error analysis

Some of the disagreements between the two annotators in the double-annotation phase are caused by ambiguities of clinical statements, sometimes erroneous narratives in notes. For instance, our annotators made different decisions on following snippets of texts: 不除外不均 (cannot exclude inequality), 不排除无压痛 (cannot exclude no tenderness). Such meaningless or misleading statements are usually caused by typos entered by physicians. Some other disagreements were caused by the presence of signals of both certainty and uncertainty, such as in following piece of text found in the discharge summary when describing process of treatment: “炎症症状明显？继续抗炎治疗” (see clear evidence of infection? Continue anti-infection therapies). The question mark was believed to be a typo given the context. Many of the disagreements, though, are simply caused by missing annotation by one of the annotators. Appendix B lists all disagreements between the two annotators which are not such mismatches of common cues due to missing annotation by one coder. The annotators then agreed to abandon such cases, making no annotations when a statement is meaningless. In our study, all disagreements between the annotators were guaranteed to be resolved before moving forward to the single annotation stage.

Our annotators originally believed that their rule set can solve most of the cases, and therefore there is little need of machine learning for the task. However, the experimental results indicate that the rules produce too many false positives, although they can indeed cover most of the cues and have roughly the same recall value with machine learning based systems. Around half of the false positives were brought by the following keywords: 略 (a little), 稍 (a little), 时有 (occasionally), since they are common Chinese characters/words with multiple meanings. Complex constraints must be added to correct such false positives, which may harm the sensitivity of the system. 时有 (occasionally), specifically, yields interesting errors regarding word segmentation, such as in ‘9时有护士查房 (the nurse mad her usual rounds at 9 am)’. Some other false positives indeed express uncertainties, but do not modify the clinical terms of interest according to our guideline. It is noteworthy that although the four types of clinical terms were identified before speculation annotation and were presented to our annotators, it is difficult to make regular expression rules to strictly define when a speculation cue is actually modifying an entity of interest. For instance, in ‘疑似 (suspect) 发现 (finding)穿孔 (perforation)’, the speculation cue ‘疑似 (suspect)’ and the clinical entity ‘穿孔 (perforation)’ is separated by another word ‘发现 (finding)’. This type of exceptions makes it impossible to accurately determine whether a cue is a modifier of a clinical term without syntactic analysis. As such, no constraints were added to the rule set to get rid of false positives when speculation cues are describing other events.

Our experimental results indicate that CRF-based sequence labeling is effective in identifying speculative cues in Chinese clinical notes. Our best system achieves performance of 92.2 measured by *F* score, which is roughly on par with their state-of-the-art English counterparts [27]. The Kappa agreement between our best system, W2V-D-bigram, and the two annotators, respectively, are 0.71 and 0.74, which are almost as high as the inter-rater agreement between the two annotators. This indicates that the system may replace human coders in identifying speculation cues in most cases, although machine learning may still not be able to deal with unseen scenarios as well as humans. Most of the hedge cues in the clinical notes are certain keywords, such as “大致 (about)”, “可能性大 (very likely)”, and “疑似 (suspicious)”. However, in rare cases with little evidence in the training data, the cues may not be detected properly (e.g., “证据不足 (lack of evidence)” in our corpus).

Many of the errors made by our system are caused by incorrect boundaries. For example, in “似 (likely) 有 (to have) 腹水 (ascites)”, our gold standard annotated “似 (likely) 有 (to have)” as a speculative cue while the system labels only “似 (likely)”. In contrast to English, both “似 (likely)” and “似有 (likely to have)” are legal words in Chinese, which means either of the annotations is acceptable in practice. One reason of this issue is that in our manual annotation, the “minimal units principle” we follow is directly taken from the English guideline, which is essentially “minimal words principle”. Cases like “似有 (likely to have)” can be avoided if “minimal character principle” is followed, which reminds us to pay more attention to language-specific adjustments in future studies when annotation guidelines for Chinese NLP tasks are inherited from English studies. Boundary detection errors can hopefully be reduced if part-of-speech and syntax are added as features, since some boundary errors like “疑似为 (is suspected to be)” (gold standard: 疑似 is suspected) are caused by including unnecessary syntactic constituents of sentences. In order to get a sense of how many boundary errors occur, we also evaluated our system using a slightly different evaluation metric, which allows for partial match of the speculative cues. There, if a gold-standard cue is a one-character cue, the predicted cue must be exactly the same one; if the gold-standard cue is multi-character, then at least 50% of the characters in the gold standard cue must also be identified to be counted as a true positive prediction. Under this partial-match evaluation, our best system, W2V-D-bigram, yields results with a 95.4 *F* score.

We also decompose the results by investigating cues modifying different types of entities. Our best system, W2V-D-bigram, yields *F* scores of 93.6, 91.5, 90.9, and 89.3, for diagnosis, symptoms, laboratory test, and treatment, respectively. The differences across different types may be a result of unbalanced sizes of samples. Also, one major difference between cues modifying diseases and cues modifying other types of entities is that almost all disease related cues indicate true speculations, such as “不除外三房心 (*cannot exclude* cor triatriatum)”, while many of other ones represent vagueness, such as in “大约 3 cm 囊肿 (*about* 3 cm cyst)”. The results suggest that it may be necessary to consider speculations and other uncertainties separately in future work.

Our experimental results also show how word segmentation affects downstream NLP tasks like speculation detection. Characters in Chinese, which are roughly equivalent to syllables, or morphemes like affixes or stems in English, can indeed express meanings by themselves, but are usually more ambiguous or vague than words in meaning. Our results demonstrate that word segmentation in clinical notes can be difficult if no annotated data provided, since existing tools trained on data of a different genre perform poorly in segmenting clinical terms. The most critical message our results conveys regarding word segmentation is that although it may be good practice to use character-level representations directly in Chinese clinical NLP, a good domain specific word segmenter can significantly strengthen bag of words and word embedding and makes it a superior choice over character-based ones.

Our experimental results also demonstrate the effectiveness of embedding models in speculation detection from Chinese clinical notes. Although the technique has been applied successfully to many NLP applications, it is the first time that both character embedding and word embedding are systematically carried out and compared in a Chinese NLP task. From a clinical NLP standpoint, our study also creates opportunities for future research in optimizing representations in clinical language processing tasks.

4.2. Limitations

There are several limitations of this study. First, our CRF word segmenter, although retrained on admission notes, is learning from an overly small training set, which included only 100 samples. A larger set of training data for word segmentation may help boost the system performance even more. Second, we did not push the system to the limit by tuning parameters or carrying out sophisticated feature engineering. In fact, we only focused on basic, standard features and embedding models, since our goal was to investigate the added value of embeddings and word segmentation in the task of speculation. It might be helpful to add common features like POS tags, to combine different features, and to cascade systems with different foci. Third, most of the data used in this study is coded by single annotator, although quality control of annotation has been carried out by doing inter-rater agreement tracking on a small portion of clinical notes. Without full double annotation on the entire data set, the gold-standard may be vulnerable to random annotation errors. The fourth major limitation of this work lies in the fact that only speculative cues are identified, while scopes of the speculations are ignored. Scope finding can be equally important because it represents which facts in context are affected by the uncertainty cues, and hence facilitate further semantic analyses over the texts. Scope finding can also be approached by sequence labeling [27] and will be part of our future work. Finally, all data in this study is from a single hospital, which is one of the most prestigious hospitals in China and which also makes the most efforts in EHR data standardizing. However, since no state-level or region-level standards for EHR data has been established and implemented in China, it is likely that there could be huge differences between terminologies used in EHR narratives by different hospitals, clinics, or even physicians. Therefore, the generalizability of our system still needs to be evaluated in order to support cross-institution applications.

5. Conclusions

Our study proposes a state-of-the-art speculation detection approach for Chinese clinical text. We experimented with bag of characters, bag of words, word embedding, and character embedding as features, and demonstrate the effectiveness of embedding models, and that word segmentation is a critical step to generate high-quality word representations for downstream information extraction applications. In particular, we suggest that a domain-dependent word segmenter can be vital to clinical NLP tasks in Chinese language.

Acknowledgments

The authors would like to thank Dr. Yonghui Wu and Dr. Hua Xu, who helped with the data preparation. This study was supported by the National Natural Science Foundation of China (NSFC) Grant # 81171426 and #81471756, and National Institute of General Medical Sciences Grant R01GM114355.

Appendix A. Annotation guideline for speculation detection

1. Task and Examples

Speculation cue in this task is defined as any words, phrases, or punctuations that bring or increase the uncertainty of the statements. Two types of uncertainties were considered in this annotation:

A. True speculations, which express that a fact, such as diagnosis or history of treatment, is not known with certainty. Particularly, question marks were sometimes used in Chinese clinical notes by doctors to express uncertain diagnosis, which should be annotated as speculation cues. Example cues include:

高血压性心脏病 可能性大 (Hypertensive heart disease, *very likely*)

不排除糖尿病眼病 (*Cannot exclude* diabetic retinopathy)

肠梗阻*?* (Intestinal obstruction*?*)

考虑脑瘤 (*suggestive* of brain tumor)

B. Vagueness in describing facts, especially numeric information.

快速性房颤病史 5 年余 (*More than* 5 years history of rapid atrial fibrillation)

体温 38.5 度 左右 (Temperature *around* 38.5 degree)

2. Detailed guideline

A. Minimal unit principle

The minimal unit that expresses uncertainty is marked. For example, in 肾功能大致正常 (renal function roughly in the normal range), 大致 (roughly) instead of 大致正常 (roughly in the normal range) should be annotated

B. A cue can be split because of the syntactic structure of Chinese. For example, in:

不排除肝性脑病可能 (*Cannot exclude the possibility* of hepatic encephalopathy)

In this case, the minimal unit principle should be followed as well: only the first part, 不除外 (cannot exclude), which alone can express the uncertainty, should be marked.

C. The annotators should pay particular attention to distinguish linguistic uncertainty and clinical uncertainty.

Only the linguistic uncertainty should be annotated. An example of clinical uncertainty is as follows:

右肺炎性病变 待查 (right lung pneumonia *yet to be examined*).

In this statement, pneumonia is an uncertain diagnosis which needs to be confirmed by further examinations. However, linguistically there is no uncertainty toward the undiagnostic fact. This can be particularly confusing in Chinese language.

D. A cue should be annotated only when it is modifying following types of information: diagnosis and diseases, signs and symptoms, laboratory tests, and treatment procedures and methods. Other cues are not considered, even if it is describing clinical events. For example:

约五点进入急诊室 (enter the ED at *about* 5 pm)

E. Do not annotate if a statement is meaningless to human, usually due to typos in the clinical notes.

Appendix B

Disagreements between users which are not caused by missing annotation by one of the coders are listed below. Underline parts of texts in the three columns are speculation cues annotated by the two annotators and in the resolved annotation, respectively.

Annotator A	Annotator B	Resolved
多处斑痕 (multiple sites of scars)	多处斑痕 (multiple sites of scars)	多处斑痕 (multiple sites of scars)
不除外不均 (cannot exclude inequality)	不除外不均 (cannot exclude inequality)	不除外不均 (cannot exclude inequality)
不除外无压痛 (cannot exclude no tenderness)	不除外无压痛 (cannot exclude no tenderness)	不除外无压痛 (cannot exclude no tenderness)
转运中可能存在的风险 (possible risks during transfer)	转运中可能存在的风险 (possible risks during transfer)	转运中可能存在的风险 (possible risks during transfer)
炎症症状明显? 继续抗炎治疗 (See clear evidence of infection? Continue anti-infection therapies)	炎症症状明显? 继续抗炎治疗 (See clear evidence of infection? Continue anti-infection therapies)	炎症症状明显? 继续抗炎治疗 (See clear evidence of infection? Continue anti-infection therapies)
未见明显异常现象 (not see clear abnormity)	未见明显异常现象 (not see clear abnormity)	未见明显异常现象 (not see clear abnormity)
数次出现 (appear multiple times)	数次出现 (appear multiple times)	数次出现 (appear multiple times)

References

[1] D. Li, K. Kipper-Schuler, G. Savova, Conditional random fields and support vector machines for disorder named entity recognition in clinical texts, Proc. Curr. Trends Biomed. Nat. Lang. Process. (BioNLP 2008) (2008).

[2] D. Demner-Fushman, W.W. Chapman, C.J. McDonald, What can natural language processing do for clinical decision support?, J Biomed. Inform. (2009) 760–772 (PMID: 19683066).

[3] M. Jiang, Y. Chen, M. Liu, S.T. Rosenbloom, S. Mani, J.C. Denny, et al., A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, J. Am. Med. Inform. Assoc. 18 (5) (2011) 601–606 (PMID: 21508414).

[4] C. Friedman, P.O. Alderson, J.H. Austin, J.J. Cimino, S.B. Johnson, A general natural-language text processor for clinical radiology, J. Am. Med. Inform. Assoc. 1 (2) (1994) 161–174 (PMID: 7719797).

[5] Y. Wu, J. Leia, W.Q. Wei, B. Tang, J.C. Denny, S.T. Rosenbloom, et al., Analyzing differences between Chinese and English clinical text: a cross-institution comparison of discharge summaries in two languages, Stud. Health Technol. Inform. 192 (2013) 662–666.

[6] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu, A comprehensive study of named entity recognition in Chinese clinical text, J. Am. Med. Inform. Assoc. 21 (Ml) (2014) 808–814.

[7] H. Wang, W. Zhang, Q. Zeng, Z. Li, K. Feng, L. Liu, Extracting important information from Chinese Operation Notes with natural language processing methods, J. Biomed. Inform. 48 (2014) 130–136 (Elsevier Inc.; PMID: 24486562).

[8] Y. Xu, Y. Wang, T. Liu, J. Liu, Y. Fan, Y. Qian, et al., Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries, J. Am. Med. Inform. Assoc. 21 (2014) e84–92 (PMID: 23934949).

[9] H.G. Chang, Z. Hai, Chinese word segmentation: a decade review, J. Chin. Inf. Process. 21 (3) (2007) 8–20.

[10] Y. Lee, K. Papineni, S. Roukos, Language model based Arabic word segmentation, ACL (2003) 399–406.

[11] R. Bar Haim, K. Sima'an, Y. Winter, Choosing an optimal architecture for segmentation and POS-tagging of Modern Hebrew, ACL Work Comput. Approaches to Semit. Lang. (June) (2005) 39–46.

[12] M. Sassano, An empirical study of active learning with support vector machines for japanese word segmentation, ACL (July) (2002) 505–512.

[13] N. Xue, F. Xia, F. Chiou, M. Palmer, The Penn Chinese TreeBank: phrase structure annotation of a large corpus, Nat. Lang. Eng. 11 (2) (2005) 207–238 (Internet).

[14] H. Tseng, P. Chang, G. Andrew, et al. A conditional random field word segmenter, in: Proc. 4th SIGHAN Work. Chinese Lang. Process., 2005 (X).

[15] R. Collbert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537.

[16] J. Turian, L. Ratinov, Y. Bengio, Word representations: a simple and general method for semi-supervised learning, in: Proc. 48th Annu. Meet. Assoc. Comput. Linguist., 2010, pp. 384–394.

[17] R. Pivovarov, N. Elhadad, A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts, J. Biomed. Inform. 45 (3) (2012) 471–481 (Internet, Elsevier Inc.; 2012 Jun, cited 2014 Mar 5).

[18] A. Henriksson, H. Dalianis, S. Kowalski, Generating features for named entity recognition by learning prototypes in semantic space: the case of de-identifying health records, in: Bioinforma Biomed. (BIBM), 2014 IEEE Int. Conf., 2014, pp. 450–457.

[19] S. Jonnalagadda, T. Cohen, S. Wu, G. Gonzalez, Enhancing clinical concept extraction with distributional semantics, J. Biomed. Inform. 45 (1) (2012) 129–140 (Elsevier Inc.).

[20] N. Elhadad, S. Zhang, P. Driscoll, S. Brody, Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions, in: Proc. AMIA Annu. Fall Symp., 2014.

[21] Y. Wu, M. Jiang, J. Lei, H. Xu, Named entity recognition in Chinese clinical text using deep neural network, Stud. Health Technol. Inform. 216 (2015) 624–628.

[22] H. Chen, Deep learning for Chinese word segmentation and POS tagging, EMNLP (October) (2013) 647–657.

[23] M. Light, X.Y. Qiu, P. Srinivasan, The language of bioscience: facts, speculations, and statements in between, in: BioLink 2004 – Proc. Work Link Biol. Lit. Ontol. Databases, 2004, pp. 17–24.

[24] V. Vincze, G. Szarvas, R. Farkas, G. Móra, J. Csirik, The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes, BMC Bioinformatics 9 (2008) 1–9.

[25] N.P. Cruz Díaz, M.J. Maña López, J.M. Vázquez, V.P. Álvarez, A machine-learning approach to negation and speculation detection in clinical texts, J. Am. Soc. Inf. Sci. Technol. 63 (7) (2012) 1398–1410.

[26] N.P.C. Díaz, Detecting Negated and Uncertain Information in Biomedical and Review Texts, RANLP, 2013, pp. 45–50.

[27] R. Farkas, V. Vincze, G. Móra, J. Csirik, G. Szarvas, The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text, in: Proc. Fourteenth Conf. Comput. Nat. Lang. Learn., 2010, pp. 1–12.

[28] E. Velldal, Predicting speculation: a simple disambiguation approach to hedge detection in biomedical literature, J. Biomed. Semantics 2 (Suppl. 5) (2013) S7 (PMID: 22166306).

[29] K. Roberts, S.M. Harabagiu, A flexible framework for deriving assertions from electronic medical records, J. Am. Med. Inform. Assoc. 18 (5) (2011) 568–573 (PMID: 21724741).

[30] S. Agarwal, H. Yu, Detecting hedge cues and their scope in biomedical text with conditional random fields, J. Biomed. Inform. 43 (6) (2010) 953–961 (Elsevier Inc.).

[31] S. Zhang, H. Zhao, G. Zhou, B. Lu, Hedge detection and scope finding by sequence labeling with procedural feature selection, Comput. Linguist. (July) (2010) 92–99.

[32] B. Tang, X. Wang, X. Wang, B. Yuan, S. Fan, A cascade method for detecting hedges and their scope in natural language text, in: Proc. Fourteenth Conf. Comput. Nat. Lang. Learn., 2010, pp. 13–17.

[33] B. Medlock, T. Briscoe, Weakly supervised learning for hedge classification in scientific literature, in: Proc. 45th Meet. Assoc. Comput. Linguist., 2007, pp. 992–999.

[34] D.A. Hanauer, Y. Liu, Q. Mei, F.J. Manion, U.J. Balis, K. Zheng, Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients, in: AMIA Annu. Symp. Proc. 2012, 2012 (September 2015), pp. 321–30 (PMID: 23304302).

[35] K. Fujikawa, K. Seki, K. Uehara, A hybrid approach to finding negated and uncertain expressions in biomedical documents, in: Proc. 2nd Int. Work Manage. interoperability Complex Heal. Syst. – Mix '12, 2012, p. 67.

[36] A. Özgür, D.R. Radev, Detecting speculations and their scopes in scientific text, in: Proc. 2009 Conf. Empir. Methods Nat. Lang. Process., 2009 August, pp. 1398–1407.

[37] V. Ganter, M. Strube, Finding hedges by chasing weasels: hedge detection using Wikipedia tags and shallow linguistic features, in: Proc. ACL-IJCNLP 2009 August, pp. 173–176.

[38] Z. Chen, B. Zou, Q. Zhu, P. Li, The scientific literature corpus for chinese negation and uncertainty identification, Chinese Lex Semant, Springer, Berlin Heidelberg, 2013, pp. 657–667.

[39] Z. Chen, B. Zou, Q. Zhu, P. Li, Chinese negation and speculation detection with conditional random fields, Nat. Lang. Process. Chinese Comput., Springer, Berlin Heidelberg, 2013, pp. 30–40.

[40] B. Zou, Q. Zhu, G. Zhou, Negation and speculation identification in Chinese language, in: Proc. 53rd Annu. Meet. Assoc. Comput. Linguist., 7th Int. Jt. Conf. Nat. Lang. Process., 2015, pp. 656–665.

[41] F. Ji, X. Qiu, X. Huang, Exploring uncertainty sentences in Chinese, in: Proc. 16th China Conf. Inf. Retrieval, 2010, pp. 594–601.

[42] Y. Xu, L. Chen, J. Wei, S. Ananiadou, Y. Fan, Y. Qian, et al., Bilingual term alignment from comparable corpora in English discharge summary and Chinese discharge summary, BMC Bioinformatics 16 (1) (2015) 1–10.

[43] J. Cohen et al., A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46 (Durham).

[44] Y. Xu, J. Tsujii, E.I.-C. Chang, Named entity recognition of follow-up and time information in 20000 radiology reports, J. Am. Med. Inform. Assoc. 19 (5) (2012) 792–799 (PMID: 22771530).

[45] L. He, Z. Yang, H. Lin, Y. Li, Drug name recognition in biomedical texts: a machine-learning-based method, Drug Discov. Today 19 (5) (2014) 610–617 (PMID: 24140287).

[46] A. Lamurias, T. Grego, F.M. Couto, Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI, BioCreative Chall. Eval. Work., 2013, (Cdi), p. 75.

[47] T. Kudo, CRF++: Yet Another CRF Toolkit, Softw. available, 2005. <http://crfpp.sourceforge.net>.

[48] P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman, Natural language processing: an introduction, J. Am. Med. Inform. Assoc. 18 (5) (2011) 544–551 (PMID: 21846786).

[49] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 16 Jan 2013, pp. 1–12, arXiv:1301.3781.

[50] P.C. Chang, M. Galley, C.D. Manning, Optimizing Chinese word segmentation for machine translation performance, in: Proc. Third Work. Stat. Mach. Transl., 2008, pp. 224–232 (June).

[51] Contemporary Chinese Dictionary. <https://en.wikipedia.org/wiki/Xiandai_Hanyu_Cidian> (Internet).