Gu *et al.*
*BMC Medical Informatics and Decision Making* (2021) 21:363
https://doi.org/10.1186/s12911-021-01716-2

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

# Composition-driven symptom phrase recognition for Chinese medical consultation corpora

Xuan Gu[1,2], Zhengya Sun[1,2*] and Wensheng Zhang[1,2]

## Abstract

**Background:** Symptom phrase recognition is essential to improve the use of unstructured medical consultation corpora for the development of automated question answering systems. A majority of previous works typically require enough manually annotated training data or as complete a symptom dictionary as possible. However, when applied to real scenarios, they will face a dilemma due to the scarcity of the annotated textual resources and the diversity of the spoken language expressions.

**Methods:** In this paper, we propose a composition-driven method to recognize the symptom phrases from Chinese medical consultation corpora without any annotations. The basic idea is to directly learn models that capture the composition, i.e., the arrangement of the symptom components (semantic units of words). We introduce an automatic annotation strategy for the standard symptom phrases which are collected from multiple data sources. In particular, we combine the position information and the interaction scores between symptom components to characterize the symptom phrases. Equipped with such models, we are allowed to robustly extract symptom phrases that are not seen before.

**Results:** Without any manual annotations, our method achieves strong positive results on symptom phrase recognition tasks. Experiments also show that our method enjoys great potential with access to plenty of corpora.

**Conclusions:** Compositionality offers a feasible solution for extracting information from unstructured free text with scarce labels.

**Keywords:** Symptom phrase recognition, Named entity recognition, Medical consultation, Composition driven

## Introduction

The high-speed development of internet is changing the habits of individuals to harvest information and obtain answers to the questions. Nowadays, more and more people try to figure out their physical problems via online consultation with medical professionals[1]. Accordingly, a great number of corpora containing the communications between patients and doctors are accumulated. However, the rare resources of doctors would be hard to satisfy the growing demands for medical services. This prompts the use of medical consultation corpora for the development of medical automatic question answering (QA) systems which have been studied over several decades [1, 2].

*Correspondence: zhengya.sun@ia.ac.cn
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
Full list of author information is available at the end of the article

[1] These websites include: https://www.haodf.com/, https://www.120ask.com/, and so on.

© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

**Table 1** Cases of extraction results by word matching

| Sentences | Word Matching | Ground Truth |
|---|---|---|
| "胸/里/长了/纤维瘤/演化/成/绝症/的/可能性/有/多/大?" ENG: "How likely will the fibroma in the chest evolve into the terminal disease?" | "胸/里/长了/纤维瘤/演化/成" | "胸里长了纤维瘤" ENG: "the fibroma in the chest" |
| "急/着/出门/气温/下降/呼吸/突然/困难，该/吃/什么/药/好？" ENG: "It was suddenly hard to breathe when I was in a hurry to go out and the temperature dropped. What medicine should I take?" | "急/下降/呼吸/突然/困难" | "呼吸困难" ENG: "hard to breathe" |
| "前/几天/不止/半夜/易/惊醒，而且/感觉/胸闷、心跳/加速。这/是/什么/原因/呢？" ENG: "I was easily awakened up at midnight the other day, feeling chest distress and rapid heartbeat. What reason is it?" | "前/不止/半夜/易/惊醒，感觉/胸闷/心跳/加速" | "半夜易惊醒，胸闷、心跳加速" ENG: "easily awakened up at midnight, chest distress and rapid heartbeat" |
| "发病/时/膝盖/以下/直接/没/知觉，有/什么/治疗/方案/吗?" ENG: "I felt no sensation below the knee when illness occurred. Is there any treatment option?" | "发病/时/膝盖/以下/直接/没/知觉" | "发病时膝盖以下没知觉" ENG: "no sensation below the knee when illness occurred" |
| "左/小腹/偶尔/有/抽筋/似的/疼痛/是/怎么/回事？" ENG: "There was occasional cramping pain in the left lower abdomen. What's the matter?" | "左/小腹/有/抽筋/疼痛" | "左小腹疼痛" ENG: "pain in the left lower abdomen" |
| "肚脐/下方/两/根/手指/处/隐疼/一/年/半/了，我/该/怎么办？" ENG: "I have felt dull pain in the lower abdomen, two fingers below the navel, for one and a half years. What should I do?" | "肚脐/下方/手指/处/隐疼" | "肚脐下方隐疼" ENG: "dull pain below the navel" |

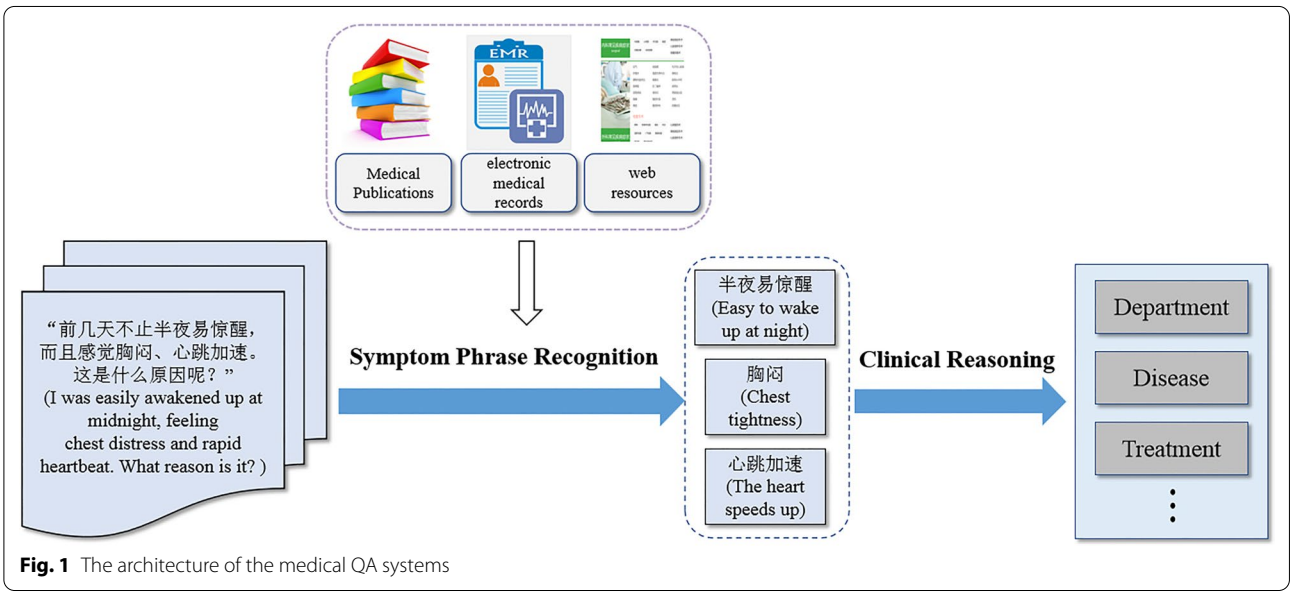


**Fig. 1** The architecture of the medical QA systems

The majority of automatic QA systems rely on named entity recognition (NER) as the first step [3, 4]. In clinical domains, named entity recognition refers to the automatic identification of text spans which represent particular entities (e.g., symptoms, diagnoses, medications) [5]. For this study, we focus on the task of symptom phrase recognition, because patient symptoms are integral to health care communications, and diagnostic and therapeutic reasoning. In particular, we concern about the statements which are stored in Chinese medical consultation corpora. Most existing methods for identifying symptoms require either enough manually annotated training data or as complete a symptom dictionary as possible. However, this is often not available in domain specific scenarios. On the one hand, the supervised medical textual resources are quite scarce, because to annotate the data sets needs the experts' knowledge and experience, which involve high overhead. On the other hand, the spoken language expressions are so diverse that

it is difficult to collect all the patients' descriptions of the symptoms.

Like humans, a medical QA system (The architecture of the medical QA systems can be depicted as in Fig. 1.) should be able to leverage known vocabulary to understand the meaning of the processed information. Suppose there is a list of standard symptom phrases used by the domain practitioners, we are interested in recognizing and extracting symptom phrases from patients' descriptions of what they are experiencing. One straightforward way for exploiting the vocabulary list is to match the entire phrases and decide what symptoms to extract. However, the word sequence repeatedly used by the domain practitioners is an ad-hoc description. Many other wordings in oral expression are used to describe subjective feelings. E.g. "心口憋着难受" (ENG: "shortness of breath and in a state of discomfort") describes the typical symptom of "胸闷" (ENG: "chest distres"). It is often the case that most symptom mentions appearing in the consultation corpus fail to match the symptom phrases in the vocabulary list. An alternative way is to rely on tokenization and match individual words to acquire symptom information. However, this may lead to either incorrect boundaries (e.g., the former three cases in Table 1) or inappropriate collocations (e.g., the latter three cases in Table 1). Note that we use the notation '/' in Chinese sentences to denote the segmentation using the tokenizer.[2]

To tackle these issues, we go beyond word matching and propose a **Com**posi-tion **D**riven (ComD) symptom phrase recognition method for Chinese medical consultation corpora without any annotations. The basic idea is to directly learn models that capture the composition, i.e., the arrangement of the symptom components. We introduce an automatic annotation strategy for the standard symptom phrases which are collected from medical publications, electronic medical records, and web resources. Specifically, we establish a position recognition model based on the relative positions between the symptom components (semantic units of words). Afterwards, we learn the embedding representations for the components, which are then used to estimate the interaction scores between them. By integrating the position outputs and the interaction scores, we are allowed to recognize the symptom phrases in medical consultation corpora. Experimental results demonstrate the feasibility and effectiveness of the proposed method which further improves the overall performance.

[2] here we use "Jieba" Chinese language segmentation module that is implemented in python https://pypi.org/project/jieba/.

The contributions of our paper can be summarized into three aspects:

- We view each symptom phrase in the vocabulary as composition of words and their interactions. This allows us to deduce the symptom phrases that are not seen before, without dependence upon any annotated corpus.
- We incorporate the position outputs and the interaction scores to judge the compositionality between individual words during symptom prediction, which can be viewed as an innovative attempt in the field of data mining.
- Experiments have shown that for symptom phrases recognition tasks, the proposed method can achieve strong positive results, and have great potential with access to increasing online textual corpus.

## Related work

For a medical question answering system, understanding the symptom phrases from the patient's input is the most critical step in providing an effective solution. A significant way to address the issues of symptom phrases recognition is named entity recognition (NER), which is the task to identify mentions of rigid designators from text belonging to predefined semantic types such as person, location, organization, etc. Over the past few decades, NER has made great strides in a wide range of areas with the help of technologies such as artificial rules, traditional machine learning (ML), and deep learning (DL). Here, we classify these existing NER techniques into three levels from the attributes of methods, and conduct a brief analysis of the pros and cons of some representative strategies related to our work below.

### Rule-based approaches

Some hand-crafted rules, for example, domain-specific gazetteers and syntactic-lexical patterns, are commonly used to design the rule-based NER systems. Kim [6] adopted Brill rule inference method for speech input, making the system generates rules automatically based on Brill's part of speech markers. In the field of biomedicine, Hanisch et al. [7] proposed ProMiner, which utilizes pre-processed thesaurus to identify protein mentions and potential genes in biomedical texts. Quimbaya et al. [8] presented a dictionary-based NER method based on electronic health records, and experimentally verified that the approach improves recall while having limited impact on precision. UMLS [9] is one biomedical resource which is prepared by medical experts manually. It has a metathesaurus which contains terms and codes from many vocabularies. Luca et al. [10] proposed

QuickUMLS: a fast, unsupervised, approximate dictionary matching algorithm for medical concept extraction. Similar approaches identify entities largely through handcrafted semantic and syntactic rules and then work very well when lexicon can be exhaustive. However, domain-specific rules and incomplete dictionaries make such systems tend to have high precision and low recall, making it impossible to transfer to other domains.

#### ML-based approaches

In machine learning, NER is cast into a multi-class classification or clustering task, which learns a model that identifies similar patterns from unseen data. Among the supervised approaches, Bikel et al. [11, 12] proposed an NER system based on Hidden Markov Models (HMM), namely IdentiFinder, to classify names, date, time expressions and numerical quantities. Besides, McCallum et al. [13] proposed a feature-induced method for Conditional Random Fields (CRFs) in NER, which achieves F-score of 84.04% for English by performing on CoNLL03. Krishnan et al. [14] presented a two-stage approach of coupling two CRF classifiers, in which the second CRF utilizes the potential representations from the output of the first CRF. Moreover, there are plenty of supervised NER strategies based on other ML algorithms, Decision Trees [15], Maximum Entropy Models [16] and Support Vector Machines (SVMs) [17] for examples, which have been studied and successfully applied by many scholars. Admittedly, supervised learning algorithms rely on a large amount of annotated data, which is time-consuming and laborious. As a result, unsupervised NER approaches are more desirable. Collins et al. [18] observed that the use of unlabeled data reduces the requirements for supervision to just seven simple seed rules, and then proposed two unsupervised algorithms for the classification of named entities. Nadeau et al. [19] presented an unsupervised gazetteer building and named entity ambiguity resolution system that combines entity extraction with disambiguation based on simple and efficient heuristics. Besides, Zhang et al. [20] proposed an unsupervised method for extracting named entities from biomedical texts by relying on terminologies, corpus statistics and shallow syntactic knowledge, and experiments on two mainstream biomedical databases proved the effectiveness and universality of the method. In the field of semi-supervised approach, Ke et al. [21] proposed using Co-training combining with CRF and SVM on Chinese organization name recognition. Co-training is a semi-supervised learning method, which uses a small amount of tagged corpus and large scales of untagged corpuses for machine learning. Liu et al. [22] proposed to combine a K-Nearest Neighbors (KNN) classifier with a linear Conditional Random Fields(CRF) model under a semi-supervised learning framework to recognize entities for tweets. Actually, the main characteristic of the ML-based approaches is to identify the combination of feature extraction and model selection that work well together for enhanced prediction performance [23–25]. In particular, they extract context features as sources of semantic encoding variability, which may have limitations in the face of less rigid and more flexible spoken language.

#### DL-based approaches

In recent years, deep learning, empowered by continuous real-valued vector representations and semantic composition through nonlinear processing, has been employed in NER systems, yielding state-of-the-art performance [26]. The application of neural models for NER was pioneered by [27], where an architecture based on temporal convolutional neural networks over word sequence was proposed. BiLSTM-CRF [28], as the most commonly-used architecture for NER using deep learning, combines BiLSTM and CRF and effectively solves the problem of handling the strong dependence of tags in the sequence ineffectively. Recently, Batbaatar et al. [29] proposed a novel neural network architecture, named semantic-affective neural network (SENN), which utilizes semantic/syntactic and emotional information by using pre-trained word representations. Besides English, there are some studies on Chinese language. Wu et al. [30] studied NER in the Chinese clinical literatures. Zhang et al. [31] proposed an LSTM model of lattice structure for Chinese NER, which encodes the sequence of input characters and all potential words matching the vocabulary. Li et al. [32] pre-trained BERT model on the Chinese clinical domain corpora, and designed a new post-processing way to combine the terminology dictionary with the model and apply radical features to the model on two Clinical Named Entity Recognition (CNER) datasets. Historically, the advantages of deep learning have been less obvious when working with small databases. For example, on the 203,621-word CoNLL-2003 English database, the best DL model, measured by F1 score, outperformed the best shallow model by only 0.4%. In other words, a large amount of annotated data is required to train a good deep learning model.

Although NER has been extensively studied in the biomedical field, symptom phrases seem to have been shelved because there is so little work being done on the subject, especially for Chinese medical texts, which we are mainly discussing here. In our work, we make full use of public resources to mine medical knowledge,

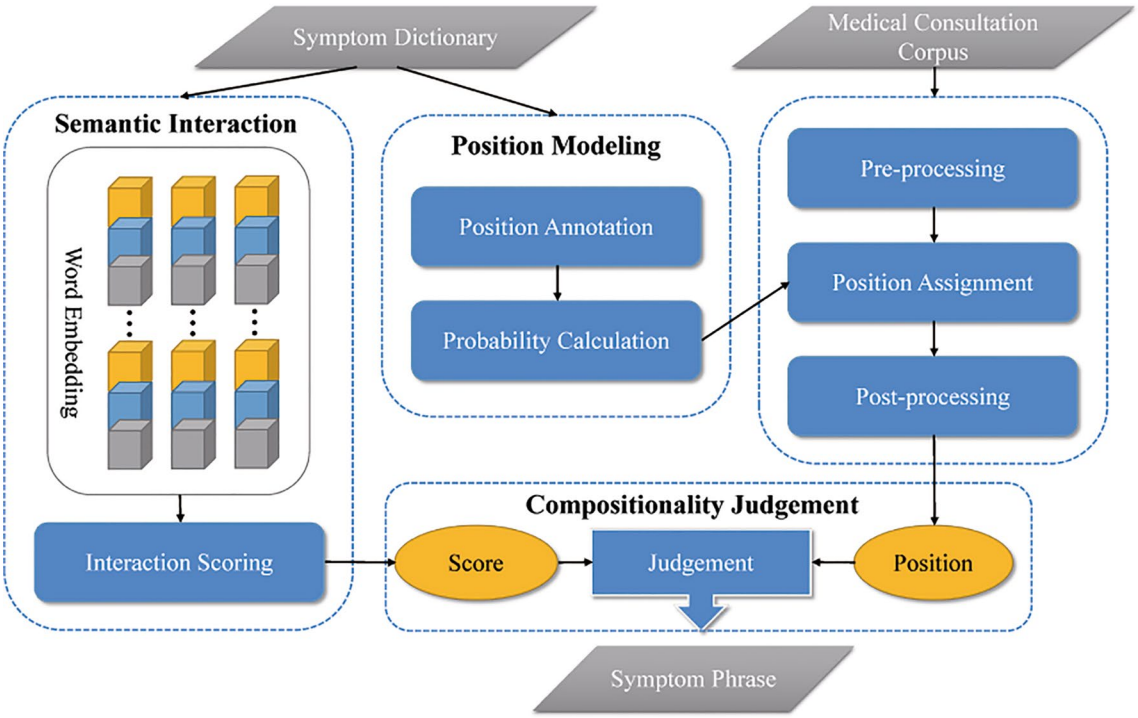


**Fig. 2** The workflow of ComD

and does not require manual annotations. By devising a novel automatic annotation strategy, the proposed approach can save a lot of labour and material costs. More specifically, we combine the position information and the interaction scores between the components of the symptom phrases, and then employ them to identify new symptom phrases in the corpus.

## Proposed method

We begin by introducing a general framework for unsupervised symptom phrase recognition. We then describe three crucial components in detail, i.e., (1) how to characterize the positions of the symptom components, (2) how to calculate the interaction scores between the symptom components, and (3) how to recognize unobserved symptom phrases.

### Overview

For this study, we build a symptom dictionary in advance by aggregating a number of standard symptom phrases from multiple resources (i.e., medical publications, electronic medical records, and web resources). Equipped with the dictionary, we are allowed to mine useful knowledge, which facilitates the downstream task of symptom extraction from medical consultation corpus.

- From the perspective of position, we provide an automatic annotation scheme to indicate demarcation of symptom phrases, followed by characterizing the probabilities of each component (semantic units of words) arranged on different positions within a symptom phrase.
- From the perspective of interaction, we learn the embedding for each component to capture the contextual associations between them. If two components tend to co-occur or appear in similar contexts, they will be mapped into similar word vectors.

When medical consultation corpus is concerned, we use a basic pipeline to pre-process the texts involved. First, we segmented each text into sentences using a Chinese text analysis tool[3]. We next divided each sentence into words by the tokenizer[4]. After that, we filtered out non-informative words such as extremely common words, rare words, and those that don't appear in the symptom dictionary. The retained words are then assigned with annotations that contribute to extracting the results. During post-processing, we identify the most probable

[3] https://github.com/blmoistawinde/HarvestText.

[4] here we use "Jieba" Chinese language segmentation module that is implemented in python https://pypi.org/project/jieba/.

**Table 2** Some basic composition forms of Chinese symptom phrases

| Composition form | Examples | Word-based annotations | Character-based annotations |
|---|---|---|---|
| Simple word | "发热"<br>*ENG:* "fever" | 发热/S | 发/B 热/E |
| Modifier + Center word | "先天性贫血"<br>*ENG:* "congenital anemia" | 先天性/B 贫血/E | 先/B 天/I 性/I 贫/I 血/E |
| Part word + Center word | "心绞痛"<br>*ENG:* "heart angina" | 心/B 绞痛/E | 心/B 绞/I 痛/E |
| Negative Word + Center word | "无疼痛"<br>*ENG:* "no pain" | 无/B 疼痛/E | 无/B 疼/I 痛/E |
| Center word + Modifier | "腹泻恶化"<br>*ENG:* "worsening diarrhea" | 腹泻/B 恶化/E | 腹/B 泻/I 恶/I 化/E |
| Modifier + Part word + Center word | "持续性腰部疼痛"<br>*ENG:* "persistent low back pain" | 持续性/B 腰部/I 疼痛/E | 持/B 续/I 性/I 腰/I 部/I 疼/I 痛/E |

boundaries of the candidate symptom phrases. Following the principle of compositionality, we integrate the position information and the interaction effect to recognize the symptom phrases that are not seen before. Figure 2 presents the workflow of ComD which comprises three main modules, i.e., position modeling (detailed in Sect. 3.2), semantic interaction (detailed in Sect. 3.3) and compositionality judgement (detailed in Sect. 3.4).

### Position modeling

Different from traditional annotation methods ([33]) which manually annotate each word in the textual corpora, in this task, we consider annotations in the symptom dictionary, which can be realized automatically. As mentioned previously, the symptom dictionary includes a list of standard expressions about patient symptoms. Through observing a variety of Chinese symptom phrases, it was found that there are mainly two types from the view of morphology, i.e., simple words (e.g. "胸闷", ENG: "chest distress") and compound words (e.g. "胸口/不通畅", ENG: "chest is not smooth"). In addition, a compound word can be split into multiple simple words by the tokenizer mentioned before. Therefore, we adopt the basic annotation signs "BIES" (Begin, Intermediate, End, Single) to represent the position information of a simple word in the phrase. For convenience, we abbreviate a "simple word" as a "word", unless otherwise stated.

Table 2 shows some basic composition forms of Chinese symptom phrases, and illustrates how the standard symptom phrases are annotated. For example, if there is only one simple word, such as "发热" (ENG: "fever"), then it is annotated as "S". If there is more than one simple word, such as "持续性/腰部/疼痛" (ENG: "persistent low back pain"), then the word "持续性" (ENG: "persistent") is annotated as "B" indicating its beginning position, the word "疼痛" (ENG: "pain") is annotated as "E" indicating its end position, and the other word "腰部" (ENG: "low back") is annotated as "I" indicating its intermediate position.

Besides word-based annotations, we can also conduct character-based annotations whose signs "BIES" represent the position information of a character in the symptom phrase, as illustrated in Table 2.

Given a component in a symptom phrase, we count how many times it appears on a particular position. For example, the word "间歇性" (ENG: "intermittent") appears 60 times at the beginning and 12 times in the middle. We model the probability of counts as a multinomial distribution $(n, \pi_B, \pi_I, \pi_E, \pi_S)$, where $\pi_B$, $\pi_I$, $\pi_E$, and $\pi_S$ denote the probabilities of four possible positions on each of $n$ independent trials. For example, in terms of "间歇性" (ENG: "intermittent"), there are 72 independent trials, each of which leads to a success for exactly one of the four positions "BIES". The parameters can be derived based on maximum likelihood estimation. In the above example, the estimations are (0.83, 0.17, 0, 0). Note that these results allow us to infer candidate symptom phrases from the perspective of position arrangements.

### Semantic interaction

We exploit the vector representation of words computed on the symptom dictionary, and implement certain metric to estimate an interaction score between any two words in the embedding space.

(a) *Word embedding* For this study, we build the embedding using the Word2Vec implementation proposed by Mikolov et al. [34], a shallow, two-layer neural network based on a skip-gram model [35]. As is well known, the skip-gram model is directed toward the prediction of the surrounding words given a target word as input (Fig. 3). The learned embedding allows to capture the

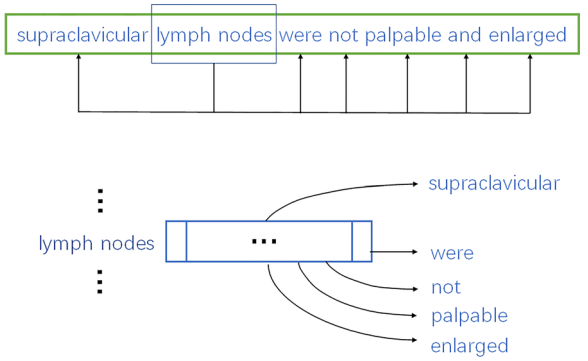


**Fig. 3** Description of the skip-gram model. The model used in Word2Vec to find an optimal representation to predict the surrounding context of a target word. Consider a standard symptom phrases from the symptom dictionary, "锁骨上淋巴结未触及肿大" (ENG: "supraclavicular lymph nodes were not palpable and enlarged"). The example highlights the window around "淋巴结" (ENG: "lymph node"), organs that produce immune cells for fighting infections. The target word, "淋巴结" (ENG: "lymph node"), is linked to each of its neighboring words and the pairs are fed into the network. The learning process optimizes the probability of predicting the contextual words of "淋巴结" (ENG: "lymph node")

contextual associations between components in a symptom phrase. In this sense, if two components tend to co-occur or appear in similar contexts, they will be mapped to approximate vectors. Note that there are several neural embedding models which are known to be fully expressive, and which may thus be thought of as more promising candidates for learning word representations [36, 37]. We address whether neural embedding models are able to capture compositions of words and their interactions in the next section.

(b) *Interaction scoring* Once the embeddings are built, we are in a position to predict interaction scores based on the distribution of word vectors in the word embedding. A straightforward way to measure proximity between two components is to use a common distance metric such as Euclidean distance, Manhattan distance or Cosine similarity, and regard them to interact with each other when the metric is within a certain threshold. However, this is problematic due to the fact that the close proximity in high dimensional space does not necessarily imply strong semantic interaction. As high dimensionality often carries rich and diverse semantics, the same distance may arise from different regions of the semantic space. Inspired by [38], we assess the degree of semantic interaction between two components according to the data distribution in their neighbourhood. Intuitively, the less similar neighbours they share, the weaker the semantic interaction is.

Following [38], we first use K-means algorithm [39] to partition the vectors into multiple clusters. Meanwhile, we apply KNN algorithm [40] to find the nearest neighbours. By analyzing the cluster membership of the nearest neighbours, we are allowed to associate each component vector with a discrete probability distribution. In this sense, the discrete probability distribution is derived from the clusters that the neighbours belong to and the corresponding occupancy. Endowed with explicit semantics, the resulting representation helps to support the calculation of interaction scores. We use KL divergence [41] to calculate how much information is lost when approximating one distribution with another. The formula is as follows:

$$D_{KL}(P_\alpha||P_\beta) = \sum_i P_\alpha(i) \log \left[\frac{P_\alpha(i)}{P_\beta(i)}\right], \quad (1)$$

where $P_\alpha$ and $P_\beta$ refer to the discrete probability distribution matrices of two components $w_\alpha$ and $w_\beta$, respectively. For example, for the words "持续性" (ENG: "persistent") and "发热" (ENG: "fever"), $P_\alpha$ represents a discrete probability distribution corresponding to "持续性" (ENG: "persistent") and $P_\beta$ represents a discrete probability distribution corresponding to "发热" (ENG: "fever").

In order to circumvent the asymmetry of KL divergence, we use a score function based on the JSD, defined as follows [38],

$$JSD(P_\alpha||P_\beta) = \frac{1}{2}D_{KL}(P_\alpha||M) + \frac{1}{2}D_{KL}(P_\beta||M), \quad (2)$$

where $M = \frac{1}{2}(P_\alpha + P_\beta)$.

In order to measure the interactions between components $w_\alpha$ and $w_\beta$ of interest, we use the scoring function whose range is within [0, 1], defined as follows [38]:

$$I(w_\alpha, w_\beta) = \exp(-\nu JSD_{\alpha,\beta} + \gamma), \quad (3)$$

Where $\nu$ and $\gamma$ are scaling and offset parameters respectively.

### Compositionality judgement

We combine the position information and the interaction scores to judge the compositionality between individual words. Suppose the words are independent of each other. Let $\pi_B^{w_\alpha}$, $\pi_I^{w_\alpha}$, $\pi_E^{w_\alpha}$, and $\pi_S^{w_\alpha}$ signify the probabilities of "B", "I", "E" and "S" positions in term of a given word $w_\alpha$. For an input sentence from patients' descriptions, we assign each $w_\alpha$ with possible annotations according to $\pi_B^{w_\alpha}$, $\pi_I^{w_\alpha}$, $\pi_E^{w_\alpha}$, and $\pi_S^{w_\alpha}$. Besides "BIES", we use "O" to represent the "Other" position, indicating that the corresponding component is independent of the extracted results. Actually, the words in a sentence may have more than one annotations. We choose such annotations as candidates, which result in non-overlapping subsequences either assigned with "B" at

**Table 3** An example of interaction scores between the internals and the boundaries

| Annotation pairs | Component pairs | Interaction scores |
|---|---|---|
| B−I pairs | "耳朵 − 开始" (ENG: "ear−at first") | 0.32 |
| | "耳朵 − 感染" (ENG: "ear−infection") | 0.56 |
| | "耳朵 − 好" (ENG: "ear−so") | 0.12 |
| I−E pairs | "开始 − 痒痛" (ENG: "at first−itchy and painful") | 0.24 |
| | "感染 − 痒痛" (ENG: "infection−itchy and painful") | 0.63 |
| | "好 − 痒痛" (ENG: "so−itchy and painful") | 0.13 |

the beginning and "E" at the end or assigned with "S". Then, the boundary scoring function $\boldsymbol{\pi}(l_1, \ldots, l_q)$ is defined as follows:

$$\boldsymbol{\pi}(l_1, \ldots, l_q) = \sum_{i=1}^{q} \boldsymbol{\pi}^{l_i}, \tag{4}$$

where $\boldsymbol{\pi}^{l_i}$ denotes the score for the subsequence $l_i$. If $l_i$ corresponds to one single component, i.e., $l_i = w_1^i$, then the score is defined as $\boldsymbol{\pi}^{l_i} = \pi_S^{w_1^i}$. Otherwise, let $l_i = w_1^i - \cdots - w_k^i$, the score is thus defined as $\boldsymbol{\pi}^{l_i} = \pi_B^{w_1^i} \cdot \pi_E^{w_k^i}$.

Note that there may be consecutive occurrences of words with annotation "B". We treat the leftmost word as the beginning. Similarly, for the consecutive occurrences of words with annotation "E", we treat the rightmost word as the end. In general, the words that can act as boundaries can also act as intermediate components (e.g. "间歇性" (ENG: "intermittent") appears in "间歇性/尿/蛋白" (ENG: "intermittent urinary protein") and "上肢/间歇性/运动/障碍" (ENG: "upper limb intermittent dyskinesia") respectively). This also makes sense from the view of morphology.

After finding the most probable boundaries from the candidates, we are required to determine which intermediate components are useful and indispensable. This is achieved by calculating the interaction scores between each intermediate component and the boundary components. Specifically, if the subsequence is annotated as B-E, its components are both kept and directly combined into a symptom phrase. Otherwise if the subsequence $w_1^j - w_2^j - \cdots - w_k^j$ is annotated as B−I−$\cdots$−E, the component $w_i^j$ assigned with "I" will be discarded if its utility value is less than a certain threshold $\delta$. Formally, the utility function is defined as follows:

$$\boldsymbol{\theta}(w_i^j) = \frac{I(w_1^j, w_i^j) + I(w_i^j, w_k^j)}{2}, \quad 1 < i < k, \tag{5}$$

where $I(w_1^j, w_i^j)$ represents the interaction score between the beginning component $w_1^j$ and the intermediate component $w_i^j$, $I(w_i^j, w_k^j)$ represents the interaction score between the intermediate component $w_i^j$ and the end component $w_k^j$, both of which are defined in equation (3).

For example, an input sentence contains a subsequence "耳朵 − 开始 − 感染 − 好 − 痒痛" (ENG: "An ear infection is so itchy and painful at first") which is annotated as "B−I−I−I−E". Let $\delta$ be set to 0.5. Then the intermediate component "开始" (ENG: "at first") and "好" (ENG: "so") are discarded due to their utility values are less than $\delta$. As a result, the extracted symptom phrase is "耳朵感染痒痛" (ENG: "An ear infection is itchy and painful"). The details are shown in Table 3.

## Experiments

In this section, we investigate the performance of the proposed method ComD on the medical consultation corpus crawled from public websites. The goals of our experiments are threefold, where one is to investigate the contribution of each module by performing the detailed ablation study, one is to examine whether ComD is able to achieve satisfactory results compared to other baseline approaches and the other is to analyze the effect of the threshold parameter $\delta$.

### Data collection

The datasets consist of the symptom dictionary for training and the medical consultation data for prediction.

*Symptom Dictionary* contains 732,855 symptom phrases which were collected from multiple data sources including medical publications[5], electronic medical records[6] and web resources[7][8]. They are the conventional/typical expressions used by the domain practitioners.

During pre-processing, we first determined whether the sources have a clear delineation of symptom phrases.

[5] https://item.jd.com/10139850.html, https://item.jd.com/12419020.html.

[6] https://www.biendata.xyz/competition/ccks_2021_clinic/data/.

[7] https://www.99.com.cn/, http://www.39.net/, https://www.120ask.com/.

[8] https://dxy.com/, https://www.guahao.com/, https://www.haodf.com/.

If so, we directly loaded them into the dictionary. Otherwise, we predefined several lexico-syntactic patterns to detect symptom phrases in texts. For instance, patterns like "The principal manifestation of $NP_x$ is $NP_y$", or "Typical symptoms of $NP_x$ include $NP_y$" often indicate symptom phrases of the form $NP_y$. The punctuation marks in them and duplicate ones are removed subsequently.

*MedConSult* is a collection of nearly one thousand medical consultation records derived from the website[9]. The symptom labels are given by the human annotators. The annotations are used as the ground truth for evaluating the overall performance of the proposed method.

*Curated Data from MedConSult* is a subset of the medical consultation data MedConSult by removing the records that have no annotation signs "I". In other words, we only kept the records that contain symptom phrases with at least three components. It has been stated that if a record only contains B—E subsequences, we directly combine the boundary components into a symptom phrase without judging their semantic interaction. Therefore, this subset serves the purpose of demonstrating the necessity of each training operation (i.e., position modeling and interaction scoring), and exploring how the recognition performance is sensitive against the variations of the interaction parameter.

### Experimental settings

This section describes the metrics that are used to quantitatively evaluate our method.

*Interaction over Union (IoU)* is the most commonly used metric for comparing the similarity between two strings. The higher the IoU, the closer the extracted result is to the ground truth. In our case, if IoU between the extracted result and the ground truth exceeds a certain threshold $\varepsilon$, the extracted result is assumed to be correct.

*Micro-averaging (Micro)* is an average metric that computes the total number of false positives (FP), false negatives (FN), and true positives (TP) over all consultation records, and then computes the precision, recall, and $F_1$ score using these counts, which are defined as follows,

$$Micro\text{-}P = \frac{TP}{TP + FP} \times 100\%, \tag{6}$$

$$Micro\text{-}R = \frac{TP}{TP + FN} \times 100\%, \tag{7}$$

$$Micro\text{-}F_1 = \frac{2 \times Micro\text{-}P \times Micro\text{-}R}{Micro\text{-}P + Micro\text{-}R} \times 100\%. \tag{8}$$

*Macro-averaging (Macro)* is an average metric that treats all records equally, no matter how many symptom phrases they contain. Specifically, it computes the precision and recall independently for each consultation record $s \in S$, and then take the average over the size of the set $S$. The results are then combined to obtain the $F_1$ score.

$$Macro\text{-}P = \frac{\sum_{s \in S} P_s}{|S|} \times 100\%, \tag{9}$$

$$Macro\text{-}R = \frac{\sum_{s \in S} R_s}{|S|} \times 100\%, \tag{10}$$

$$Macro\text{-}F_1 = \frac{2 \times Macro\text{-}P \times Macro\text{-}R}{Macro\text{-}P + Macro\text{-}R} \times 100\%, \tag{11}$$

where $P_s$ and $R_s$ denote the precision and recall of the consultation record $s$ respectively.

#### *Baseline methods*

We consider such a setting that only a symptom dictionary including numerous standard symptom phrases is available, while there is no annotated dataset for clinically motivated symptom extraction. In this new setting, we compare our proposed method ComD against the well-developed dictionary-based method and the deep learning method which have achieved the current state-of-the-art on the NER datasets.

- ComD: We leverage the symptom dictionary to learn the arrangements of the symptom components. By incorporating the position outputs and the interaction scores, we are allowed to deduce the symptom phrases that are not seen before.
- BERT-CRF [42]: This method obtains its token representation from the pre-trained BERT model, which is then fed into CRF output layer for token-level classification over the NER label set. As the model requires annotated dataset, for a fair comparison, we use the symptom dictionary to retrieve the relevant consultation records and make annotations accordingly.
- BiLSTM-CRF [43]: This is a character-based CNN-BiLSTM-CRF method for Chinese named entity recognition, which enhances Chinese character representations by character glyphs. The annotated dataset used for BiLSTM-CRF is the same as that used in BiLSTM-CRF mentioned above.
- BDMM [44]: This is a commonly used word segmentation method based on the given dictionary,

[9] https://www.haodf.com/.

**Table 4** Ablation study results on curated data from MedConSult

| $\varepsilon$ | Method | Macro/% | | | Micro/% | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| 0.6 | ComD | 77.82 | 78.17 | 77.99 | 77.62 | 78.17 | 77.89 |
| | ComD-NoInt | 61.62 | 61.97 | 61.80 | 61.54 | 61.97 | 61.75 |
| | ComD-NoPos | 16.20 | 16.20 | 16.20 | 16.20 | 16.20 | 16.20 |
| 0.7 | ComD | 75.00 | 75.35 | 75.18 | 74.83 | 75.35 | 75.09 |
| | ComD-NoInt | 51.06 | 51.41 | 51.23 | 51.05 | 51.41 | 51.23 |
| | ComD-NoPos | 9.15 | 9.15 | 9.15 | 9.15 | 9.15 | 9.15 |
| 0.8 | ComD | 61.62 | 61.97 | 61.80 | 61.54 | 61.97 | 61.75 |
| | ComD-NoInt | 26.41 | 26.76 | 26.58 | 26.57 | 26.76 | 26.67 |
| | ComD-NoPos | 5.63 | 5.63 | 5.63 | 5.63 | 5.63 | 5.63 |
| 1.0 | ComD | 60.92 | 61.27 | 61.09 | 60.84 | 61.27 | 61.05 |
| | ComD-NoInt | 15.14 | 15.49 | 15.31 | 15.38 | 15.49 | 15.44 |
| | ComD-NoPos | 2.11 | 2.11 | 2.11 | 2.11 | 2.11 | 2.11 |

which combines positive maximal matching and reverse maximal matching algorithm.

- Dictionary-based: This is a dictionary lookup method that relies heavily on exact string matching, where the words, and order of words should be exactly the same as the entry in the symptom dictionary.

We implemented ComD in python. To investigate the principle of compositionality, we simplify ComD to the case in which characters are taken as symptom components. We call this method ComD-Character.

#### *Parameter settings*

We performed a grid search to tune hyperparameter values for ComD. Each of these parameters is varied from low to high at a fixed interval, and the performance on the medical consultation data is measured. Through trial and error tuning in our implementation, the following choices of hyperparameters are preferred: We set the dimension of word embeddings to 500. We used K-means with 500 clusters, and identified the $k$-nearest neighbours, with $k = 2000$. We set the scaling parameter $\nu$, offset parameter $\gamma$ and the utility threshold $\delta$ to 7.5, 0 and 0.2 respectively.

### Ablation study

To investigate the individual contribution of our position modeling and semantic interaction in training, we removed them to offer the methods ComD-NoPos and ComD-NoInt. Without position modeling, the components that interact with each other and appear in a sentence are extracted as a symptom phrase. Without semantic interaction, the components within the boundaries are directly combined into a symptom phrase.

Table 4 reports the macro/micro-averaged results on curated data from MedConSult for different IoU thresholds $\varepsilon$ from 0.6 to 1.0, all shown in percentage. As can be seen, the joint framework shows apparently superior performance in terms of macro/micro-average precision, macro/micro-average recall, and macro/micro-average $F_1$ score. For example, ComD outperforms ComD-NoInt by more than 16% when $\varepsilon = 0.6$, and this improvement is even more pronounced when increasing the parameter $\varepsilon$. This affirms the necessity of semantic interaction, and its ability to retain useful and indispensable components within the boundaries. Compared with ComD-NoPos, it can be observed that ComD increases its performance by an even larger margin. This suggests that position modeling gives more influence to the quality of extractions from unstructured text in spoken form.

### Overall comparison

We now evaluate the performance of the proposed method and other baselines on MedConSult. The detailed results are summarized in Table 5. As can be seen, our method, ComD, empirically leads to significant gains on recall and $F_1$ metrics, with the precision values slightly lower than that of BERT-CRF. Although BERT-CRF is able to achieve the highest precision values, it is not able to retrieve many symptom phrases shown by the low recall values, only a little better than that of dictionary-based methods. We believe that the good performance of ComD is due to an appropriate design of the model according to the compositionality. As a counterpart, ComD-Character does not perform as well as ComD, albeit its performance superior to the other

**Table 5** Performance comparison of the proposed and baseline methods on the MedConSult

| $\varepsilon$ | Method | Macro/% | | | Micro/% | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| 0.6 | ComD | 41.02 | **37.92** | **39.41** | 39.01 | **21.84** | **28.00** |
| | ComD-Character | 36.23 | 33.37 | 34.74 | 34.13 | 13.62 | 19.47 |
| | BERT-CRF | **47.17** | 16.57 | 24.52 | **47.16** | 15.32 | 23.13 |
| | BiLSTM-CRF | 36.50 | 29.58 | 32.68 | 34.01 | 8.43 | 13.51 |
| | BDMM-based | 8.23 | 12.15 | 9.81 | 5.92 | 8.58 | 7.00 |
| | Dictionary-based | 13.52 | 7.72 | 9.83 | 13.76 | 7.72 | 9.89 |
| 0.7 | ComD | 33.53 | **30.08** | **31.71** | 31.11 | **17.42** | **22.33** |
| | ComD-Character | 28.36 | 25.76 | 27.00 | 26.30 | 10.49 | 15.00 |
| | BERT-CRF | **37.56** | 13.05 | 19.38 | **37.44** | 12.03 | 18.21 |
| | BiLSTM-CRF | 32.00 | 25.80 | 28.57 | 26.18 | 7.51 | 11.67 |
| | BDMM-based | 7.14 | 10.46 | 8.49 | 5.10 | 7.37 | 6.02 |
| | Dictionary-based | 9.51 | 5.69 | 7.12 | 9.48 | 5.38 | 6.86 |
| 0.8 | ComD | 27.45 | **24.26** | **25.75** | 25.23 | **14.12** | **18.11** |
| | ComD-Character | 21.81 | 19.44 | 20.55 | 19.78 | 7.89 | 11.28 |
| | BERT-CRF | **27.52** | 9.48 | 14.10 | **27.45** | 8.82 | 13.35 |
| | BiLSTM-CRF | 19.25 | 14.94 | 16.82 | 14.99 | 3.61 | 5.82 |
| | BDMM-based | 6.17 | 8.77 | 7.25 | 4.22 | 6.15 | 5.00 |
| | Dictionary-based | 8.65 | 5.14 | 6.44 | 8.72 | 4.94 | 6.31 |
| 1.0 | ComD | 25.05 | **22.62** | **23.77** | 23.37 | **13.08** | **16.78** |
| | ComD-Character | 19.42 | 17.26 | 18.28 | 17.61 | 7.03 | 10.04 |
| | BERT-CRF | **26.56** | 9.11 | 13.57 | **26.55** | 8.53 | 12.91 |
| | BiLSTM-CRF | 15.00 | 12.52 | 13.65 | 12.59 | 3.03 | 4.89 |
| | BDMM-based | 5.94 | 8.54 | 7.01 | 4.10 | 5.98 | 4.86 |
| | Dictionary-based | 8.65 | 5.13 | 6.44 | 8.72 | 4.94 | 6.31 |

The best result with bold font for each parameter/model/characteristic

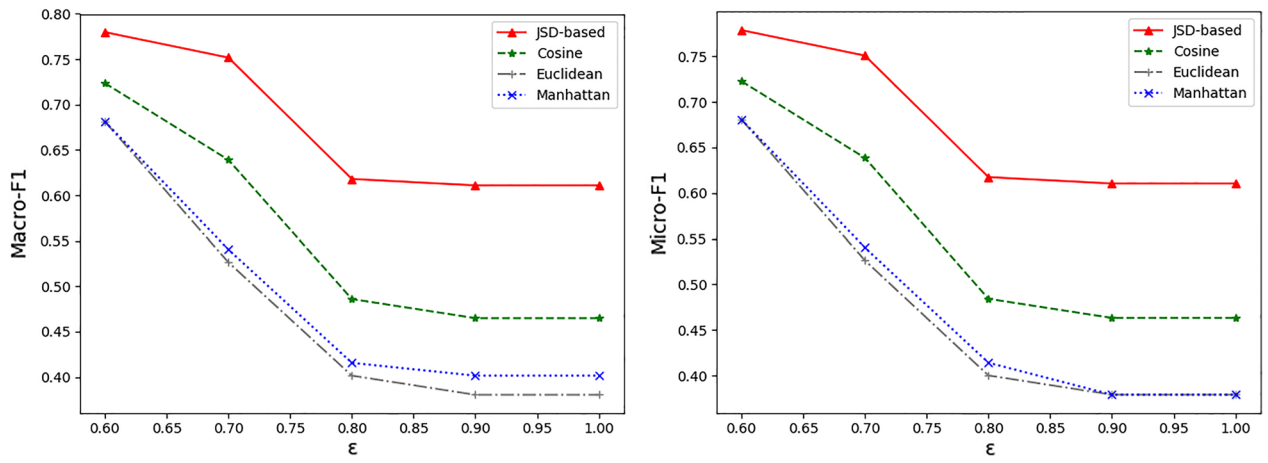


**Fig. 4** Performance comparison between JSD-based and other distance measures

methods. This is because the semantic units of characters can have multiple meanings in different words, and tend to be confusing indicators for symptom boundaries. When one compares performance of ComD and BiLSTM-CRF, macro-average precision of BiLSTM-CRF at $\varepsilon = 0.7$ appears to be comparative, but the others are relatively poor. BiLSTM-CRF has shown promise in learning Chinese character embeddings. However, it suffers from the training issue: we use symptom dictionary to provide distant supervision so that it could not exploit its

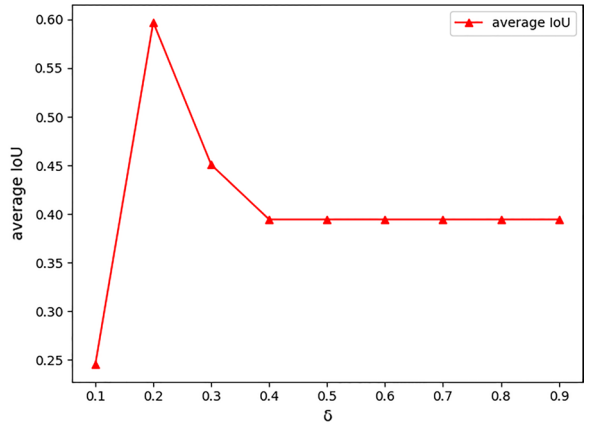


**Fig. 5** $\delta$ sensitivity comparison

full capabilities when facing flexible and variable expressions in the consultation corpus. As expected, BDMM and dictionary lookup methods are inferior to machine learning algorithms. The poor results can be explained by the fact that dictionary based methods are incapable of generalizing beyond the provided word sequences and this limitation is unlikely to be fully compensated by better matching techniques (e.g., BDMM). This also suggests that our method effectively utilizes the composition to detect more symptoms in the consultation corpus which were not present in the symptom dictionary.

In addition to the JSD-based scores presented in Sect. 3.3, we also try other distance measures for interaction scoring. Figure 4 plots the curves on macro/micro-average F1 score versus IoU thresholds $\varepsilon$. We find that the results decrease quickly within a certain interval (In this case from 0.6 to 0.85), and then the variations remain slim. This agrees with intuition that a large IoU threshold requires a high match between the extracted symptom phrases and the ground truth. Besides, the larger the value of $\varepsilon$ is, the wider the performance gaps between JSD-based and other distance measures are. This sheds light on the strength of JSD-based metric in capturing the semantic interaction in high dimensional space.

**Table 6** Typical Case

| Cases | Ground truth | Extraction result |
|---|---|---|
| S1. “女儿/体弱多病/怎么办？”<br>ENG: “My daughter is physically weak. What should I do?” | “体弱多病”<br>ENG: “physically weak” | No |
| S2. “我/的/脚趾甲/往肉/里长/怎么/治？”<br>ENG: “How should my ingrown toenails be treated?” | “脚趾甲往肉里长”<br>ENG: “ingrown toenails” | No |
| S3. “最近/几天/腰累/酸疼，是/什么/原因？”<br>ENG: “What was the reason for my waist pain and soreness in recent days?” | “腰累酸疼”<br>ENG: “waist pain and soreness” | No |
| S4. “孩子/右侧/先天性/唇鄂/裂开/并且/延伸/至/喉咙”<br>ENG: My child has the congenital cleft lip and palate on the right side, which has extended to the throat | “右侧先天性唇鄂裂开至喉咙”<br>ENG: “congenital cleft lip and palate on the right side extended to the throat” | “右侧先天性唇鄂裂开”<br>ENG: “congenital cleft lip and palate on the right side |
| S5. “我/头痛/流/鼻水，该/怎么/治疗”<br>ENG: “I am suffering from headache and a snotty nose. What should I do?” | “头痛流鼻水”<br>ENG: “headache and a snotty nose” | “头痛”<br>ENG: “headache” |
| S6. “上/牙齿/掉块/是/何原因？”<br>ENG: “Why did pieces of upper teeth fall out?” | “上牙齿掉块”<br>ENG: “pieces of upper teeth fall out” | “牙齿掉块”<br>ENG: “pieces of teeth fall out” |
| S7. “眼睑/长/了/东西, 是/个/小/疙瘩”<br>ENG: “Something growing on the eyelid is a small pimple.” | “眼睑长疙瘩”<br>ENG: “a pimple growing on the eyelid” | “眼睑长东西疙瘩”<br>ENG: “something growing on the eyelid is a pimple” |
| S8. “脚部/常年/感到/冰冷/且/时常/伴有/ 疼痛, 是/何原因？”<br>ENG: “What is the reason for the icy cold feet all the year round, and sometimes accompanied by pain?” | “脚部冰冷疼痛”<br>ENG: “icy cold feet & pain” | Yes |
| S9. “耳根/上/有时/会起/一片/大大小小/不规则/的/红点/怎么办？”<br>ENG: “What to do with a mass of irregular red spots over the root of my ear?” | “耳根会起红点”<br>ENG: “red spots over the root of my ear” | Yes |

### Sensitivity analysis

We examine how the extraction performance is sensitive against parameter variations. Figure 5 plots the curves on average IoU versus the threshold parameter $\delta$. We observe that the average IoU has a sharp rise when $\delta$ exceeds 0.1, while reaching its peak at 0.2. This can be attributed to the increased ratio of indispensable components needed for symptom recognition. It is noteworthy that when $\delta$ exceeds 0.2, the extraction quality drops quickly, and then seems to remain stable beyond a certain value (in this case around 0.4). This suggests that relatively large $\delta$ would incur performance loss until all the indispensable intermediate components are discarded.

### Case study

In this section, we analyze several representative examples to illustrate the advantages and disadvantages of the proposed ComD, as shown in Table 6. Each case consists of three columns, the first column being the selected cases, the second column being the ground truth, and the third column being the extracted results.

- Sentence S1 makes a claim about a patient experiencing a symptom "体弱多病" (ENG: "physically weak"). In fact, the phrase "体弱多病" (ENG: "physically weak") is a Chinese idiom which expresses a certain denotation as a whole, and cannot be split into multiple words by the tokenizer. As such, the proposed model fails to recognize the symptom phrase.

- Sentence S2 describes a phenomenon "脚趾甲往肉里长" (ENG: "ingrown toenails") experienced by a patient, which is further segmented into three words, i.e., 脚趾甲/往肉/里长 (ENG: "ingrown toenails"). In our case, they are out-of-vocabulary (OOV) words, and have no word embedding representation learned. Hence, they are ignored when the proposed model is applied.

- Sentence S3 makes a claim about a patient experiencing the symptom "腰累酸疼" (ENG: "waist pain and soreness"), which composes of two words, i.e., "腰累" (ENG: "waist pain") and "酸疼" (ENG: "soreness"). Here, "腰累" (ENG: "waist pain") is an OOV word and ignored during extraction. Although "酸疼" (ENG: "soreness") is a valid component of symptom phrases, it rarely appears at the starting position. Without detecting appropriate starting boundaries, the model outputs no extraction results.

- Sentence S4 contains two subsequences relevant to patient symptoms, with the primary symptom being "右侧/先天性/唇鄂/裂开/至/喉咙" (ENG: "right side of the congenital cleft lip and palate extends to the throat"). The subsequences are annotated as 右侧/B 先天性/B 唇鄂/I 裂开/E至/I and 喉咙/B (ENG: "right side of the congenital cleft lip and palate extends to the throat"). For this case, the proposed model chooses the most probable boundary, i.e., 右侧/B (ENG: "right side") and 裂开/E (ENG: "cleft"), and leave out the annotated component "至" (ENG: "to") and "喉咙" (ENG: "throat") beyond the boundaries.

- Sentence S5 makes a claim about a patient experiencing the symptoms "头痛流鼻水" (ENG: "headache and a snotty nose"). However, our method only identifies and extracts "头痛" (ENG: "headache") , leaving "流鼻水" (ENG: "a snotty nose") unrecognized. This is because the two words appearing in the phrase "流/鼻水" (ENG: "a snotty nose") are OOV words, and ignored accordingly.

- Sentence S6 is segmented into five words, i.e., "上/牙齿/掉块/是/何原因？" (ENG: "Why did small pieces of upper teeth fall out?"), where the former three relate to the symptoms. Through annotating, "上" (ENG: "upper") and "牙齿" (ENG: "teeth") are both considered as candidate boundaries at the beginning. The proposed model discards the word "上" (ENG: "upper") by mistake according to the estimated boundary scores.

- Sentence S7 describes a phenomenon "眼睑长疙瘩" (ENG: "a pimple growing on the eyelid") observed by a patient. After finding the most probable boundaries, i.e., 眼睑/B 长/I 东西/I 疙瘩/E (ENG: "Something growing on the eyelid is a pimple"), the proposed model calculates the interaction scores between each intermediate component and the boundaries, and then keeps both intermediate components, one of which is yet semantically redundant and unnecessary.

- Sentence S8 and S9 are positive examples that show our model can correctly recognize the symptom phrases. As can be seen, there are disjoint words present in a sentence, which are integral parts of a symptom mention. This is a common phenomenon in colloquial expressions. For example, the words "脚部" (ENG: "feet"), "冰冷" (ENG: "icy cold") and "疼痛"

(ENG: "pain") in sentence S8 are disjoint but are all key components of the symptom "脚部冰冷疼痛" (ENG: "icy cold feet & pain"), similar cases in sentence S9. Our model captures such composition and concatenates them as the extraction result.

## Conclusions

In this paper, we explore how using symptom dictionary can facilitate identifying symptom phrases from medical consultation corpus. The basic idea is to learn models for semantic compositionality over linguistic units according to the observed symptom phrases. Our method can not only support computer-assisted diagnosis systems, but can also promote the medical knowledge graph construction. Experimental results prove the superiority of our method. A special emphasis for our future work is placed on more effective design patterns of composing words/characters to form sound symptom expressions. We believe that compositionality provides a feasible solution for extracting information from unstructured free text with scarce labels. Another focus concerns development of computer-assisted medical consultation systems integrated with the proposed symptom recognition method.

**Acknowledgements**
We would like to thank Siheng Zhang and Liangchen Hu for helpful discussions.

**Authors' contributions**
ZS and XG conceived the idea and conducted the analyses. XG collected the data, implemented the experiments and wrote the initial draft of the paper. WZ contributed to refining the ideas and carrying out additional analyses. All authors discussed the results and revised the manuscript.

**Funding**
This research work was supported by the National Natural Science Foundation of China (No. 61876183) and the Natural Science Foundation of Beijing Municipality (No. 4172063).

**Availability of data and materials**
The datasets generated and analysed during the current study are available from the corresponding author upon reasonable request.

**Declaration**

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]University of Chinese Academy of Sciences, Beijing, China. [2]Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Received: 1 August 2021 Accepted: 8 December 2021
Published online: 27 December 2021

**References**

1. Vijeta BV. A restricted domain medical question answering system. Int J Sci Res. 2014;3(5):1602–5.
2. Abacha A, Zweigenbaum P. Means: a medical question-answering system combining nlp techniques and semantic web technologies. Inf Process Manag. 2015;51:570–94.
3. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc. 2011;18(5):580–7.
4. Chen S, Argentinis E, Weber G. Ibm Watson: how cognitive computing can be applied to big data challenges in life sciences research. Clin Ther. 2016;38(4):688–701.
5. Steinkamp JM, Bala W, Sharma A, Kantrowitz JJ. Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. J Biomed Inform. 2020;102:1–9.
6. Kim JH, Woodland PC. A rule-based named entity recognition system for speech input; 2000.
7. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J. Prominer: rule-based Protein and gene entity recognition. Bmc Bioinform. 2005;6(Suppl 1): S14.
8. Quimbaya AP, Múnera AS, Rivera RAG, Rodríguez JCD, noz Velandia OMM, Pe na AAG, Labbé C. Named entity recognition over electronic health records through a combined dictionary-based approach. Proc Comput Sci. 2016;100:55–61.
9. Bodenreider O. The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Res. 2004;32:267–70.
10. Luca S, Nazli G. Quickumls: a fast, unsupervised approach for medical concept extraction. In: 39th ACM international conference on research and development in information retrieval (SIGIR 2016); 2016.
11. Bikel DM, Miller S, Schwartz R, Weischedel R. Nymble: a high-performance learning name-finder. In: Proceedings of the fifth conference on applied natural language processing; 1997. pp. 194–201.
12. Bikel DM, Schwartz R, Weischedel RM. An algorithm that learns whats in a name. Mach Learn. 1999;34(1):211–31.
13. McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: CONLL '03 proceedings of the seventh conference on natural language learning at HLT-NAACL; 2003. vol. 4, pp. 188–191.
14. Krishnan V, Manning CD. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics; 2006, pp. 1121–1128.
15. Szarvas G, Farkas R, Kocsor A. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In: DS'06 proceedings of the 9th international conference on discovery science; 2006. pp. 267–278.
16. Borthwick A, Sterling J, Agichtein E, Grishman R. Nyu: description of the mene named entity system as used in muc-7, MUC; 1998.
17. McNamee P, Mayfield J. Entity extraction without language-specific resources. In: COLING-02 proceedings of the 6th conference on Natural language learning; 2002. vol. 20, pp. 1–4.
18. Collins M, Singer Y. Unsupervised models for named entity classification. In: 1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora; 1999.
19. Nadeau D, Turney PD, Matwin S. Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity. Lecture notes in computer science; 2006. pp. 266–277.
20. Zhang S, Elhadad N. Unsupervised biomedical named entity recognition. J Biomed Inform. 2013;46(6):1088–98.
21. Ke X, Li S. Chinese organization name recognition based on co-training algorithm. In: 2008 3rd International conference on intelligent system and knowledge engineering; 2008. vol. 1, pp. 771–777. https://doi.org/10.1109/ISKE.2008.4731034.
22. Liu X, Zhang S, Wei F, Zhou M. Recognizing named entities in tweets; 2011. pp. 359–367.
23. Tuncer T, Dogan S, Akbal E. A novel local senary pattern based epilepsy diagnosis system using eeg signals. Aust Phys Eng Sci Med. 2019;42:939–48.

24. Tuncer T, Ertam F. Neighborhood component analysis and relieff based survival recognition methods for hepatocellular carcinoma. Phys A Stat Mech Appl. 2020;540:123143.
25. Tuncer V, Dogan S, Ertam F, Subasi A. A novel ensemble local graph structure based feature extraction network for eeg signal analysis. Biomed Signal Process Control. 2020;61:102006.
26. Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. IEEE Trans Knowl Data Eng (2018).
27. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res. 2011;12(76):2493–537.
28. Huang Z, Xu W, Yu K. Bidirectional lstm-crf models for sequence tagging; 2015. arXiv preprint arXiv:1508.01991.
29. Batbaatar E, Li M, Ryu K. Semantic-emotion neural network for emotion recognition from text. IEEE Access. 2019;7:111866–78.
30. Wu Y, Jiang M, Lei J, Xu H. Named entity recognition in Chinese clinical text using deep neural network. Stud Health Technol Inform. 2015;216:624–8.
31. Zhang Y, Yang J. Chinese ner using lattice lstm. In: Proceedings of the 56th annual meeting of the association for computational linguistics; 2018. vol. 1, pp. 1554–1564.
32. Li X, Zhang H, Zhou X. Chinese clinical named entity recognition with variant neural structures based on bert methods. J Biomed Inform. 2020;107:103422. https://doi.org/10.1016/j.jbi.2020.103422.
33. Zheng S, Wang F, Bao H, Hao Y, Zhou P, Xu B. Joint extraction of entities and relations based on a novel tagging scheme; 2017. pp. 1227–1236. arXiv preprint arXiv:1706.05075.
34. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems; 2013. pp. 3111–3119.
35. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the international conference on learning representations; 2013. pp. 1–12.
36. Dong W, Wu J, Bai Z, Hu Y, Li W, Qiao W, Woźniak M. Mobilegcn applied to low-dimensional node feature learning. Pattern Recogn. 2021;112:107788.
37. Bai Z, Li Y, Woźniak M, Zhou M, Li D. Decomvqanet: decomposing visual question answering deep network via tensor decomposition and regression. Pattern Recogn. 2021;110:107538.
38. Matteo M, Roland M, Joris C, María R. Context-specific interaction networks from vector representation of words. Nat Mach Intell. 2019;2019:181–90.
39. Karlekar A, Seal A, Krejcar O, Gonzalo-Martin C. Fuzzy k-means using non-linear s-distance. IEEE Access. 2019;7:55121–31.
40. Hezil H, Djemili R, Bourouba H. Signature recognition using binary features and knn. Int J Biometric. 2018;10(1):1–15.
41. Huan Z, Pengzhou Z, Zeyang G. K-means text dynamic clustering algorithm based on kl divergence. In: 2018 IEEE/ACIS 17th international conference on computer and information science (ICIS); 2018. pp. 659–663.
42. Mao J, Liu W. Hadoken: a bert-crf model for medical document anonymization. In: Proceedings of the Iberian languages evaluation forum co-located with 35th conference of the Spanish society for natural language processing; 2019. pp. 720–726.
43. Yaozong J, Xiaobin X. Chinese named entity recognition based on cnn-bilstm-crf. In: 2018 IEEE 9th international conference on software engineering and service science (ICSESS); 2018. pp. 1–4.
44. Gai R, Gao F, Duan L, Sun X, Li H. Bidirectional maximal matching word segmentation algorithm with rules. Adv Mater Res. 2014;926–930:3368–72.


## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

**Learn more** biomedcentral.com/submissions