2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)

Conditional Relationship Extraction for Diseases and Symptoms by a Web Search-based Approach

1st Yi-Hui Lee*
Department of Computer Science
The University of Texas at Dallas
Richardson, Texas
yi-hui.lee@utdallas.edu

2nd Jia-Ling Koh
Department of Computer Science and Information Engineering
National Taiwan Normal University
Taipei, Taiwan
jlkoh@csie.ntnu.edu.tw

***Abstract*—This paper studies the strategies of automatically extracting the conditional relationships between diseases and symptoms from a Chinese encyclopedia site and the disease-related web pages searched from the Internet. At first, the seed symptoms of a disease are extracted from an online medical encyclopedia automatically. These seed symptoms are utilized as query keywords to automatically find more symptoms of a disease from the unstructured documents of the disease-related search results. Next, a jointly learning method is used to construct the embedded representations of the conditional terms and pattern terms. Besides, the semantic similarity matrix of conditional terms is computed through the co-clustering algorithm to discover the representative conditional terms of the clusters. The result of experiments shows that the proposed method, which discovers the semantically related symptoms of a disease associated with conditionals, achieves high accuracy. Besides, many unusually known symptoms considered by the medical experts are discovered, which may be noticeable symptoms needing further verification in the future.**

***Index Terms*—Text Mining, Information Extraction, Semantic Networks**

I. Introduction

In recent years, medical data mining has become an important research issue. More and more different data mining tasks in healthcare are studied. Most of the analytics in healthcare today focus on structured data, such as the Electronic Health Records (EHRs) are used to assist doctors in the decision making of treatments [1]. Furthermore, the track on Clinical Decision Support in the Text REtrieval Conference (TREC-CDS) [2] aimed to retrieve the diagnosis, order, and treatment from medical records to answer the medical questions. On the other hand, the unstructured healthcare data collected from the public posting on the social media platforms can be hard to manage but provide possibly useful information. For example, [3] studied how to predict the drug reactions from discussion forums, [4] detected adverse drug events in Tweets with semi-supervised convolutional neural networks, and [5] created a catalogue of real-world treatments from online Autism communities.

To build a knowledge base from the unstructured medical data extremely improves the usage and understanding of the healthcare documents. A knowledge base not only can provide answers automatically for the Question Answering (QA) systems [6]–[8] but also help semantics understanding for natural language processing [9]. A lot of tasks have been proposed to build general-purpose knowledge bases from different corpora automatically. For example, the online encyclopedia is considered to be a good resource to construct a knowledge base, such as the DBpedia[1] constructed from Wikipedia[2] [10] and the NELL [11]–[13] learned knowledge from the web.

In order to automatically process and analyze the Electronic Medical Records (EMRs) to provide a clinical decision support system, it is a critical step to retrieve the relationships between medical concepts, especially for diseases and symptoms. However, most medical knowledge bases are in English. [14] provided an approach to extract symptoms and symptom-related entities from healthcare websites and encyclopedia sites for constructing a medical knowledge base in Chinese. However, various symptoms of a disease occur under different conditions. For each disease, it is necessary to discover the specific conditional terms for the corresponding symptoms. For example, `chest pain` is a symptom of `lung cancer` in the `early stage`. Then the `has_symptom` relationship between `lung cancer` and `chest pain` should have the corresponding conditional term `early`. Accordingly, in this paper, we aimed to automatically extract the conditional relationship between diseases and symptoms from an encyclopedia site and healthcare websites in Chinese. By giving a disease, say `lung cancer`, the symptoms of `lung cancer` will be extracted associated with the conditions as triples, such as `has_symptom (lung cancer, early, chest pain)`.

The challenges of this task are as follows:

- The information provided in the Chinese medical encyclopedia is limited. Only a few symptoms of a disease can be extracted.
- It is not trivial to extract the symptoms related to a given disease from the healthcare web pages because the information is unstructured.
- It is difficult to decide the useful conditions from the contexts of a symptom.

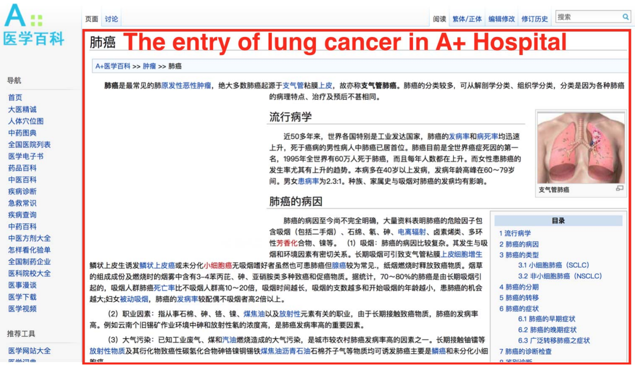
*Work performed while this author was at NTNU.

[1] http://wiki.dbpedia.org

[2] https://www.wikipedia.org

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.00-38

IEEE computer society



肺癌 The entry of lung cancer in A+ Hospital

A+医学百科 >> 肿瘤 >> 肺癌

肺癌是最常见的肺原发性恶性肿瘤，绝大多数肺癌起源于支气管粘膜上皮，故亦称**支气管肺癌**。肺癌的分类较多，可从解剖学分类、组织学分类，分类是因为各种肺癌的病理特点、治疗及预后不甚相同。

流行病学

近50多年来，世界各国特别是工业发达国家，肺癌的发病率和病死率均迅速上升，死于癌病的男性病人中肺癌已居首位。肺癌目前是全世界癌症死因的第一名，1995年全世界有60万人死于肺癌，而且每年人数都在上升。而女性患肺癌的发生率尤其有上升的趋势。本病多在40岁以上发病，发病年龄高峰在60～79岁间。男女患病率为2.3:1。种族、家属史与吸烟对肺癌的发病均有影响。

肺癌的病因

肺癌的病因至今尚不完全明确，大量资料表明肺癌的危险因子包含吸烟（包括二手烟）、石绵、氡、砷、电离辐射、卤素烯类、多环性芳香化合物、镍等。 （1）吸烟：肺癌的病因比较复杂。其发生与吸烟和环境因素有密切关系。长期吸烟可引致支气管粘膜上皮细胞增生鳞状上皮生诱发鳞状上皮癌或未分化小细胞癌无吸烟嗜好者虽然也可患肺癌但腺癌较为常见.，纸烟燃烧时释放致癌物质。烟草的组成成份及燃烧时的烟雾中含有3−4苯丙芘、砷、亚硝胺类多种致癌和促癌物质。据统计，70～80%的肺癌是由长期吸烟引起的，吸烟人群肺癌死亡率比不吸烟人群高10～20倍，吸烟时间越长，吸烟的支数越多和开始吸烟的年龄越小，患肺癌的机会越大;妇女被动吸烟，肺癌的发病率较配偶不吸烟者高2倍以上。

（2）职业因素：指从事石棉、砷、铬、镍、煤焦油以及放射性元素有关的职业，由于长期接触致癌物质，肺癌的发病率高。例如云南个旧锡矿作业环境中砷和放射性氡的浓度高，是肺癌发病率高的重要因素。

（3）大气污染：已知工业废气、煤和汽油燃烧造成的大气污染，是城市较农村肺癌发病率高的因素之一。长期接触铀镭等放射性物质及其衍化物致癌性碳氢化合物砷铬镍铜锡铁煤焦油沥青石油石棉芥子气等物质均可诱发肺癌主要是鳞癌和未分化小细

目录

1 流行病学
2 肺癌的病因
3 肺癌的类型
3.1 小细胞肺癌（SCLC）
3.2 非小细胞肺癌（NSCLC）
4 肺癌的分期
5 肺癌的转移
6 肺癌的症状
6.1 肺癌的早期症状
6.2 肺癌的晚期症状
6.3 广泛转移肺癌之症状
7 肺癌的诊断检查

Fig. 1. The webpage of lung cancer in the A+ Hospital.

Inspired by the idea of constructing DBpedia from Wikipedia, the first step of our work is to discover the relationships between diseases and symptoms according to the hyperlinks in the web pages of the A+ hospital[3]. The A+ hospital is a Wikipedia-like website in Chinese, which provides lots of healthcare information of diseases and symptoms as shown in Fig. 1. Next, to complete the symptoms of a given disease, we utilized the web search results, where many web pages provide healthcare information of diseases. How to extract possible conditional symptoms from the noisy web search results is the main issue studied in this paper.

We provide an off-line conditional relationship constructing system composed the following three modules: (1) Seed symptoms extraction: data in A+ hospital was used to initially construct the relationship between a disease and its seed symptoms, as shown in Table I. (2) Extended symptoms discovery: the disease and the seed symptoms are used as query keywords to perform web search for finding the other possible symptoms of the disease. The scoring method is proposed to select the top k candidate symptoms to form the extended symptoms as shown in Table I. (3) Conditional term discovery: the contexts of the symptom terms in the search results are collected as the candidates of conditional terms. A jointly learning approach is performed to construct the embedded representation for the conditional terms and pattern terms. Then the K-means plus plus algorithm is used to cluster the candidate conditional terms according to the similarity measure of their embedded representations. Finally, the most frequent terms in each cluster are chosen as the representative conditional terms to generate the has_symptom relationships with conditions.

The contributions of this work are summarized as follows:

- We provided a measuring method to rank the candidate symptoms of a disease extracted from the unstructured documents of the disease-related search results.
- An embedded representation learning method for each candidate conditional term was designed in an unsupervised approach.
- We applied the co-clustering approach to discover the representative conditional terms.
- We provided a framework to construct the conditional relationship of disease and symptom in Chinese.

TABLE I
THE HAS_SYMPTOM RELATIONSHIP FOR LUNG CANCER AND THE SYMPTOMS.

Seed Symptom	Extended Symptoms
胸痛(chest pain)	脱水(dehydration)
盗汗(night sweating)	谵妄(delirium)
寒战(shivering)	痴呆(dementia)
头痛(headache)	脓痰(purulent sputum)
意识障碍(disturbance of consciousness)	大小便失禁(incontinence)
瘫痪(paralysis)	视物模糊(blurring of vision)
低热(low fever)	气促(anhelation)
咳血(hemoptysis)	肝转移(liver metastases)
压痛(tenderness)	溃疡(ulcer)
心悸(palpitation)	干呕(retching)
抽搐(tic disorder)	黑蒙(amaurosis fugax)
and 15 other mores	and 89 other mores

The rest of the paper is organized as follows. After a summary of related works in Section II, we introduce the proposed methods for extracting the relationship between a disease and the symptoms in Section III. In Section IV, we present the strategies for conditional terms discovery. Section V describes the evaluation of the proposed methods. Section VI concludes this paper and gives the future work.

II. Related Work

Many data mining tasks on the medical domain were studied, such as Sun et al. [1] developed a framework, which automatically recommends treatments to doctors. Feldman et al. [3] proposed an approach to predict adverse drug reactions prior to the Food and Drug Administration (FDA). This paper proposed a text mining methodology from four online medical message boards to construct the drug-symptom relationships. Then, the lift measure is utilized to evaluate the correlation between drugs and adverse drug reactions.

Furthermore, building a knowledge base from the unstructured medical data extremely improves the understanding of the medical records. Goodwin et al. [6] presented a novel framework to answer medical questions by retrieving relevant medical articles, which is a challenge of TREC-CDS [2]. By retrieving the diagnosis, orders, and treatments from medical records, the proposed framework built a probabilistic knowledge graph: a clinical picture and therapy graph from a large collection of EMRs. Then the probabilistic inference strategy was applied to identify the answers for selecting and ranking the scientific articles containing the answers. For solving the same problem, [15] used MetaMap, a medical concept recognizer, to extract medical concepts. Besides, a Wikipedia knowledge base was used to predict the patient diagnosis. Accordingly, the original query is expanded with the predicted diagnosis to search relevant articles.

Many researches have proposed an impressive result in building general-purpose knowledge bases automatically, such as NELL [11]–[13] learning knowledge from the web by bootstrapping strategy and DBpedia constructed from Wikipedia [10]. NELL used a pattern-based strategy to build a knowledge base. The DBpedia project built a large-scale, multilingual knowledge base by extracting structured data from Wikipedia [10]. On the other hand, Wang et al. [16] extracted the

[3] http://www.a-hospital.com

concepts and the instances from Hudong, which is a Chinese encyclopedia. A method was proposed to learn ontology from the category system and Infobox schema in Hudong. Based on the ontology, the instances were extracted accordingly. Furthermore, Li et al. [17] built a cross-lingual knowledge base to integrate four wikis of different languages.

Although many methods have been developed to build knowledge bases automatically from the encyclopedia, it is possible that the information provided in the encyclopedia is not complete. For solving this problem, Savenkov and Agichtein [7] provided a Text2KB system to translate a natural language question to the Knowledge Base (KB) entities and predicates. The system utilized textual data from web search results, community question answering platforms, and a general text document collection. The topic entities in a question were detected and the question phrases were mapped to predicates in knowledge bases. West et al. [8] built an end-to-end pipeline for knowledge base completion based on search-based question answering. They used a question-answering system to retrieve relevant and up-to-date text passages in order to extract the candidate answers linking to the Freebase entities. Zhang et al. [14] constructed a knowledge base of symptoms automatically from eight healthcare websites, three Chinese encyclopedia sites, and symptoms extracted from EMRs. The categories of encyclopedia sites were used to extract target entities to train a classifier for deciding entity types. Besides, the duplications and inconsistencies in different resources were considered.

Wang et al. [18] showed that extracting the condition of a question is useful to solve the question answering problems based on a knowledge base. To extract the conditional knowledge from the dialogues, the condition terms are extracted by a bootstrapped pattern-learning method. Then the condition embedding model and the pattern embedding model are built by a supervised learning paradigm. Moreover, a new objective function is designed to modify the skip-gram model to the jointly embedding model. After that, the word embedding of conditions and patterns are utilized for co-clustering and discover the representative conditions. This paper provided us the innovative idea that condition is significant when describing a relationship between a disease and its symptoms.

III. Relation Extraction

A. Relation Extraction Problem

In our task, there are two kinds of input sources for extracting the relationships between diseases and symptoms: one is the website of A+ hospital and the other one is the non-structured documents of web search results. The goal is to find the triple *(d, c, s)*, which denotes a disease *d* has symptom *s* under the condition *c*.

The proposed approach consists of three parts of processing modules as shown in Fig. 2: (1) Seed symptoms extraction, (2) Extended Symptoms discovery, and (3) Conditional terms discovery. The details of the modules are explained in Section III-B, III-C, and Section IV, respectively.

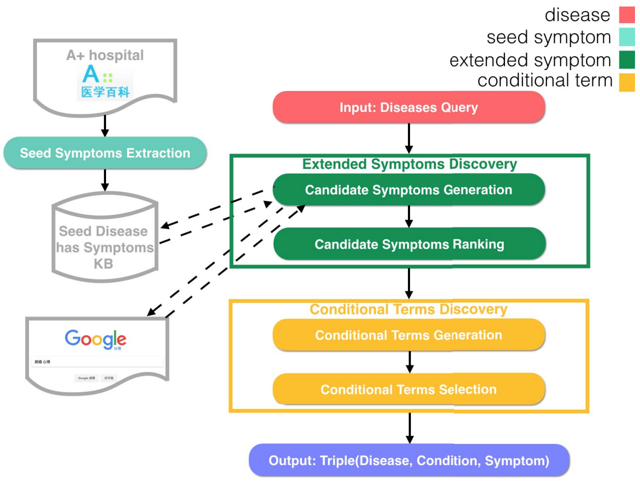
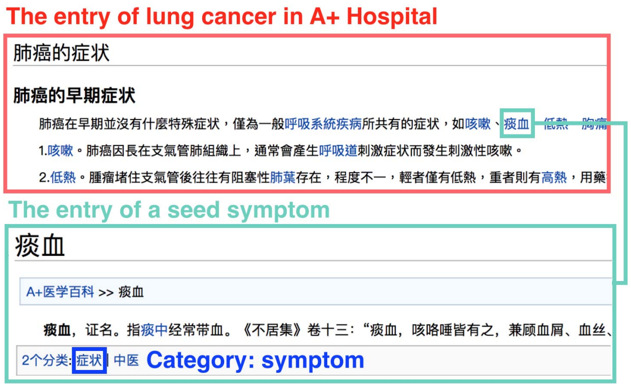

disease 🟥
seed symptom 🟩
extended symptom 🟩
conditional term 🟨
A+ hospital
A+
医学百科
Seed Symptoms Extraction
Seed Disease has Symptoms KB
Google
Input: Diseases Query
Extended Symptoms Discovery
Candidate Symptoms Generation
Candidate Symptoms Ranking
Conditional Terms Discovery
Conditional Terms Generation
Conditional Terms Selection
Output: Triple(Disease, Condition, Symptom)

Fig. 2. The system architecture.


The entry of lung cancer in A+ Hospital

肺癌的症状

肺癌的早期症状

肺癌在早期並沒有什麼特殊症状，僅為一般呼吸系統疾病所共有的症状，如咳嗽、痰血、低熱、胸痛

1.咳嗽。肺癌因長在支氣管肺組織上，通常會產生呼吸道刺激症状而發生刺激性咳嗽。

2.低熱。腫瘤堵住支氣管後往往有阻塞性肺葉存在，程度不一，輕者僅有低熱，重者則有高熱，用藥

The entry of a seed symptom

痰血

A+医学百科 >> 痰血

痰血，证名。指痰中经常带血。《不居集》卷十三：“痰血，咳咯唾皆有之，兼顾血屑、血丝、

2个分类: 症状 | 中医

Fig. 3. An example of seed symptoms extraction.

B. Seed Symptoms Extraction

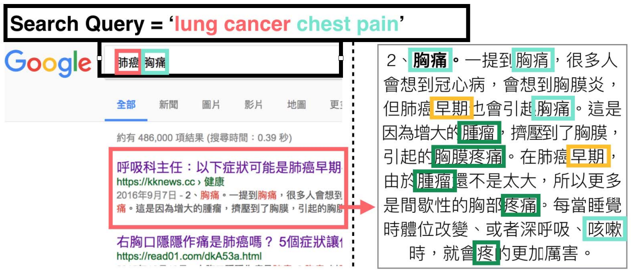
In order to extract the seed symptoms of diseases from the A+ hospital automatically, the scrapy[4] toolkit was applied to scrape all the web pages describing the diseases and symptoms. Then a Chinese dictionary of the diseases and symptoms is constructed accordingly. Besides, in the entry of a disease *d*, the mentioned terms with a hyperlink to the category `symptom` are selected to be the seed symptoms of *d* as shown in Fig. 3.

C. Extended Symptoms Discovery

A disease and its every seed symptom are used as the query keywords to search related web pages from the Internet. According to the web pages returned from the search engine, the paragraphs which contain the term 症状, which is the term of `symptom` in Chinese, are retrieved for further processing.

For example, when the disease is 肺癌`(lung cancer)` and the seed symptom is 胸痛`(chest pain)`, the query keywords given to the search engine are 肺癌`(lung cancer)` and 胸痛`(chest pain)`. One of the retrieved paragraphs is shown in Fig. 4. The terms marked in light green are seed symptoms. Besides, the other terms marked in dark green are the candidate symptoms, which are in the dictionary of symptoms.

[4] https://scrapy.org



Search Query = 'lung cancer chest pain'

約有 486,000 項結果 (搜尋時間：0.39 秒)

呼吸科主任：以下症狀可能是肺癌早期

https://kknews.cc › 健康

2016年9月7日 - 2、胸痛。一提到胸痛，很多人會想到

痛。這是因為增大的腫瘤，擠壓到了胸膜，引起的胸

右胸口隱隱作痛是肺癌嗎？ 5個症狀讓

https://read01.com/dkA53a.html

2、胸痛。一提到胸痛，很多人會想到冠心病，會想到胸膜炎，但肺癌早期也會引起胸痛。這是因為增大的腫瘤，擠壓到了胸膜，引起的胸膜疼痛。在肺癌早期，由於腫瘤還不是太大，所以更多是間歇性的胸部疼痛。每當睡覺時體位改變、或者深呼吸、咳嗽時，就會疼的更加厲害。

Fig. 4. An example of the paragraph in a search result.

In order to score the candidate symptoms, the proposed idea is that the more a seed symptom is important to the disease and a candidate symptom is related to the seed symptom, the more possible that the candidate symptom is highly related to the disease. Accordingly, the scoring function of a candidate symptom C_i consists of two parts: (1) the significance of a seed symptom S_j, i.e. $ImportantScore(S_j)$, and (2) the related weight of the candidate symptom C_i to the seed symptom S_j, i.e. $RelatedWeight(C_i, S_j)$. The defined equation is shown as below:

$$Score(C_i)=\sum_{S_j \in Seed} RelatedWeight(C_i, S_j) \times ImportantScore(S_j) \quad (1)$$

Three different strategies are designed to compute the important score of a seed symptom. The first twos use a graph model to represent the relationship among the seed symptoms, where the edges are weighted with two different methods. Then the random walk with restart paradigm [19] is used to evaluate the significant score of the seed symptoms. The third one evaluates the centrality of a seed symptom according to the average similarity measure with the other seed symptoms. The details of the three strategies are described as follows:

- Random walk with co-occurred times as edge weight (RW_C): A graph representing the relationships among the seed symptoms is constructed, where the vertices correspond to the seed symptoms. Besides, the edge weight between each pair of symptoms is assigned the co-occurred frequency when both symptoms appear in the same paragraphs of the search results within a window size of 10. After performing the PageRank algorithm [20] to determine and estimate how important each seed symptom is, which is called the representative scores. The representative scores of the seed symptoms are normalized by the maximum representative scores to get $ImportantScore(S_j)$ for each seed symptom S_j as shown in Table II. In order to score the candidate symptoms based on their relatedness with the high representative symptoms, only the seed symptoms whose important score higher than the given threshold value 0.1 are selected into $Seed$.
- Random walk with W2V cosine similarity as edge weight RW_{W2V}: According to the graph modeling the relationships among the seed symptoms, the PageRank algorithm is used to compute $ImportantScore(S_j)$ for each seed symptom S_j. Here the edge weight between two symptoms is set to be the cosine similarity of the word embedding of the symptoms. By collecting the disease-related web pages as the training corpus, the word embedding for the seed symptoms are learned from the word2vector (W2V) skip-gram model [21][5], which are provided in the gensim[6] toolkit, with window size = 10 and dimension = 250. The seed symptoms whose important scores higher than the given threshold value 0.1 are selected into $Seed$.
- Word to vector cosine similarity Average $W2V_{avg}$: According to the learned word embedding of the seed symptoms, for each seed symptom S_j, in the $W2V_{avg}$ method, $ImportantScore(S_j)$ is the average cosine similarity of S_j with the other seed symptoms. Besides, $Seed$ remains all the seed symptoms.

TABLE II
AN EXAMPLE SHOWING THE OBTAINED IMPORTANT SCORES OF THE SEED SYMPTOMS AFTER NORMALIZATION.

(score)	***s1***: **(chest pain)**	***s2***: **(cough)**	***s3***: **(malnutrition)**	***s4***: **(hemoptysis)**
Representative	1.117	1.0	0.95	0.933
Important	1.0	0.8953	0.8505	0.8353

Furthermore, the relative weight of a candidate symptom C_i to seed symptom S_j, denoted by $RelatedWeight(C_i, S_j)$, is computed by the following two methods.

- Conditional probability (CP): $RelatedWeight(C_i, S_j)$ is set to be the conditional probability $P(C_i|S_j)$, which is computed by dividing the co-occurrence frequency of C_i and S_j to the frequency of S_j in the search results.
- Word to vector semantic cosine similarity $W2V_{sim}$: $RelatedWeight(C_i, S_j)$ is set to be the cosine similarity $cos_{sim}(C_i, S_j)$, where C_i and S_j are represented by their word embedding by word2vec.

Among the candidate symptoms, the symptoms with the top *k* $Score(C_i)$ values are selected further to find their conditional terms.

IV. Conditional Term Discovery

In the conditional term discovery module, the processing consists of the following two steps: conditional terms generation and conditional terms selection.

A. Conditional Terms Generation

Based on the best performance approach in the extended symptoms discovery stage (which will be carefully discussed in Section V-B), the top *k* (*k*=100) representative symptoms are selected. The conditional terms are the ones appearing nearby the symptoms in the disease-related web pages. Therefore, the contexts appearing within a window size of 10 with the representative symptoms in the search results are collected to form the set of candidate conditional terms. In order to remove the noisy terms, only the following three types of conditional terms are remaining:

- Temporal terms: the terms whose POS tagging[7] are about time, which are denoted by C_t.

[5] Skip-gram is a model architecture to compute continuous vector representations of words from a very large data set. It tries to maximize the predicting probability of words within a certain range before and after the target word.

[6] https://radimrehurek.com/gensim/models/word2vec.html

[7] POS tags (Part-of-Speech tags) are special labels assigned to each token (word) in a sentence to indicate their syntax properties, where the POS tags t and td are about time, the POS tags adv., adj., v., and n. are about status.

- Status terms: the terms whose POS tagging are about status, which are denoted by C_s.
- Organ terms: the terms defined in the organ' s pages of the A+ hospital, which are denoted by C_o.

In the following, these three different types of candidate conditional terms are processed separately.

B. Conditional Terms Selection

In order to discover the groups of conditional terms with similar semantics, the embedded representation for each candidate conditional term is unsupervised learned firstly. For each type of conditional terms, says C_t, the co-occurred conditional terms belonging to the other two types of conditional terms, C_s and C_o, are called their pattern terms. It is assumed that the semantics of a conditional term is related to the co-occurred symptoms, conditional terms, and the pattern terms in the same paragraph. Accordingly, we apply the skip-gram model on the set of created documents, in which each contains the symptoms, conditional terms, and the patterns terms appearing together in the same disease-related search result. For example, for the conditional terms in C_t, a created document in the training data consists of the mentioned symptoms: `chest pain` and `cough`; the co-occurred conditional terms in C_t: `often` and `early`; and the pattern terms: `close to` and `heart`.

Next, the co-clustering algorithm is performed to generate the similarity matrix between candidate conditional terms. Three matrices are obtained by computing the similarity of the embedded representations between pairs of conditional terms, a conditional term and a pattern term, and pairs of pattern terms. These matrices are called the Condition-Condition similarity matrix denoted as SC with size $m \times m$, the Condition-Pattern similarity matrix demoted as M with size $n \times m$, and the Pattern-Pattern similarity matrix denoted as SR with size $n \times n$, respectively. In the matrix M, m_{ij} corresponds to the cosine similarity between the embedded representations of the i^{th} pattern term and the j^{th} conditional term, respectively.

It is assumed that the more two conditional terms are related to the semantically related pattern terms, the more these two conditional terms are related. Similarly, the more two pattern terms are related to the semantically related conditional terms, the more these two pattern terms are related. Accordingly, the Condition-Condition cosine similarity matrix is updated by the co-clustering approach iteratively.

The embedding based co-clustering is performed based on the algorithm described in Bisson et al. [22] and Wang et al. [18]. At each iteration *t*, the new similarity matrix SR_t is computed by using the similarity matrix SC_{t-1} previously computed, and so is SC_t. The defined equations are as follows:

$$SR_t = \alpha_1 M SC_{t-1} M^{\mathrm{T}} \cdot NR + (1-\alpha_1) SR_0, nr_{i,j} = \frac{1}{|m_i| \cdot |m_j|} \tag{2}$$

$$SC_t = \alpha_2 M^{\mathrm{T}} SR_{t-1} M \cdot NC + (1-\alpha_2) SC_0, nc_{i,j} = \frac{1}{|m_i| \cdot |m_j|} \tag{3}$$

The SC_0 and SR_0 represents the initial similarity matrices computed between the embedded representations for each pair of the conditional terms and pattern terms, respectively. Besides, the matrix NR and NC are used for normalization. The parameters α_1 and α_2 are used to adjust the weight to sum up the derived similarity and the initial similarity, which are set to be 0.1 in our implementations.

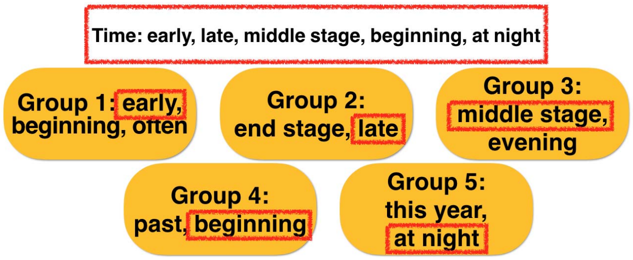
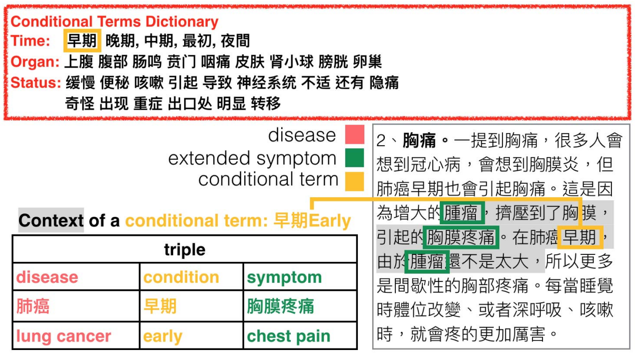

Time: early, late, middle stage, beginning, at night
Group 1: early, beginning, often
Group 2: end stage, late
Group 3: middle stage, evening
Group 4: past, beginning
Group 5: this year, at night

Fig. 5. Group the conditional terms by K-means plus plus algorithm.



Conditional Terms Dictionary
Time: 早期 晚期, 中期, 最初, 夜間
Organ: 上腹 腹部 肠鸣 贲门 咽痛 皮肤 肾小球 膀胱 卵巢
Status: 缓慢 便秘 咳嗽 引起 导致 神经系统 不适 还有 隐痛
奇怪 出现 重症 出口处 明显 转移

triple		
disease	condition	symptom
肺癌	早期	胸膜疼痛
lung cancer	early	chest pain

2、**胸痛**。一提到胸痛，很多人會想到冠心病，會想到胸膜炎，但肺癌早期也會引起胸痛。這是因為增大的腫瘤，擠壓到了胸膜，引起的胸膜疼痛。在肺癌早期，由於腫瘤還不是太大，所以更多是間歇性的胸部疼痛。每當睡覺時體位改變、或者深呼吸、咳嗽時，就會疼的更加厲害。

Fig. 6. Extract the conditional relationship triples according to the discovered conditional terms.

According to the Condition-Condition similarity matrix obtained by the co-clustering approach, the conditional terms are then clustered into *k* groups by performing the K-means Plus Plus algorithm. For example, among the candidate conditional terms with temporal type, the semantically related conditional terms 早期`(early)`, 初期`(beginning)`, 往往`(often)` are grouped into the same cluster as shown in Fig. 5

Finally, the conditional terms in each cluster with the highest frequency in the disease-related web pages are chosen to be the representative conditional terms. Therefore, in Fig. 5, the representative conditional terms of the 5 clusters are 早期`(early)`, 晚期`(late)`, 中期`(middle stage)`, 最初`(beginning)`, and 夜`(at night)`, respectively.

For a disease *d*, the discovered temporal, status, and organ representative conditional terms form its conditional terms dictionary. Finally, the conditional terms dictionary as shown in Fig. 6 is used to extract the triple relationships of a disease *d* having a symptom *s* with a condition *c*, denoted as `has_symptom`(*d*, *c*, *s*). For example, a discovered triple is `has_symptom(`肺癌`(lung cancer),` 早期 `(early),` 胸膜疼痛`(pleural pain))`.

V. Performance Evaluation

The experiments include two parts: (1) evaluation of the discovered symptoms, and (2) evaluation of the discovered conditional terms.

A. Data Description

There are 6 diseases selected to be the testing diseases: 肺癌`(lung cancer)` , 鼻咽癌`(nasopharyngeal carcinoma)`, 糖尿病`(diabetes)`, 肝硬化`(cirrhosis)`,

TABLE III
THE COLLECTED DATASET FOR EVALUATION.

	lung cancer	nasopharyngeal carcinoma	diabetes	cirrhosis	colorectal cancer	rectal cancer	in total	avg
$\#seeds$	27	7	11	13	12	15	85	14.2
$\#pages_{total}$	4272	995	1861	1962	1787	2256	13133	2188.8
$\#pages_{avg}$	158.2	142.1	169.2	150.9	148.9	150.4	919.8	153.3

大 腸 癌(colorectal cancer), and 直 腸 癌(rectal cancer), which are in the top 10 causes of death in Taiwan. Table III shows the number of seed symptoms, the total number and the average number of the disease-related web pages in search results of the seed symptoms for each disease.

B. Evaluation of the discovered symptoms

For each disease, the discovered symptoms are ranked according to their scoring results of the function $Score(C_i)$ defined in equation (1). This experiment evaluates the Mean Average Precision (MAP) of the discovered symptoms by combining the different methods to compute the important score and related weight used in the equation, respectively.

Important score: (1) RW_C, (2) RW_{W2V}, and (3)$W2V_{avg}$.
Related weight: (1) CP and (2) $W2V_{sim}$.

Given the descriptions retrieved from the disease-related web page, which contain both the disease and the discovered symptom. Each discovered symptom was labeled by non-medical experts as 0/1 according to whether the descriptions semantically imply the disease having the symptom. According to the labeled results, the macro average MAPs of the discovered symptoms across diseases were evaluated as shown in Fig. 7 Left. On the other hand, the medical experts are asked to label the discovered symptoms of the diseases score 1 if the disease usually has that symptom, score 0.5 if the disease sometimes has that symptom, and score 0 if the disease seldom has that symptom. According to the labeled scores, a discovered symptom of a disease is judged correct if its score is 1 or 0.5. The macro average MAPs are shown in Fig. 7 Right.

Fig. 7 Left shows that, overall, $RW_{W2V} + CP$ has the best performance, whose macro average MAPs across diseases achieve up to 0.85 and keep stable around 0.8. From MAP@1 to MAP@6, $RW_C + W2V_{sim}$ has the best performance, whose MAP@6 achieves up to 0.9 and 0.85 evaluated by the non-medical and medical experts, respectively. It means that this method can correctly detect the well-known top symptoms of diseases. Moreover, for the three scoring methods of important score, to combine with the CP method for computing the related weight performs better than combined with the $W2V_{sim}$ method from MAP@12 to MAP@100.

According to the results shown in Fig. 7 Right, although the MAP values evaluated by the medical experts are lower than the ones evaluated by the non-experts, their glowing curves have the similar trend. $RW_C + W2V_{sim}$ has the best performance until MAP@35, then catch up by the $W2V_{avg} + CP$. The reason that the MAP evaluated by the experts is lower than the non-expert MAP is discussed as follows. Firstly, some symptoms are too general that the experts don’t count it to be the symptoms of the disease. In the case of “There are many patients with colorectal cancer, especially those with colon cancer, found to have a certain degree of anemia in the time of the discovery of the tumor. Anemia can be manifested as dizziness, weakness, cold, dry skin, a headache, insomnia, memory loss, palpitation, shortness of breath, loss of appetite, and gastrointestinal disorders.” The non-experts labeled a headache as the symptom of colorectal cancer, while the experts thought a headache is a general symptom that may occur in many diseases. Moreover, some symptoms are caused by the metastasis of the disease. The experts determined that these symptoms are not the symptoms of the disease because the symptoms appear unusually. For example, in the case of “Metastasis causes difficulty in sucking. Hepatic metastases cause hepatomegaly and jaundice or skeletal metastases cause limbs feel sore and so on. After an examination, the symptom is caused by colorectal cancer.” That is why he non-experts labeled the hepatomegaly is a symptom of colorectal cancer according to the context but the experts didn’t. The proposed method is helpful to construct complete relationships between diseases and symptoms, which provides the noticeable symptoms for further verification in the future.

In Table IV, we compared the MAP@100 for the different diseases. It is interesting that when the disease is lung cancer, using the related weight CP has better performance than $W2V_{sim}$. However, when the disease become diabetes, using the related weight $W2V_{sim}$ performs better than using CP. It may because that there are only 11 seed symptoms of diabetes but 27 seed symptoms of lung cancer. More seed symptoms will get more related web pages as the search results. In a sparser dataset, the semantic similarity measure between a pair of candidate symptoms can show their semantic relatedness more effective than computing their conditional probability.

C. Conditional Terms Evaluation

Based on the 100 symptoms discovered by $W2V_{avg} + CP$ scoring method, the corresponding conditional terms are evaluated. The discovered conditional terms are labeled as score 0/1 according to whether the conditional term helps to understand the symptom of a disease more clearly. Accordingly, the precisions of the discovered conditional terms are computed.

In this experiment, we compared the precisions of (1) only using word embedding to compute the similarity of pairs of conditional terms for clustering and (2) the result got by the additional co-clustering step. Furthermore, two baseline methods: Baseline 1 and Baseline 2, are proposed, which choose the top *k* frequent terms from the candidate conditional terms directly. Baseline 1 set *k* equal to the number of conditional terms discovered by the word embedding method and Baseline 2 set k equal to the number of conditional terms

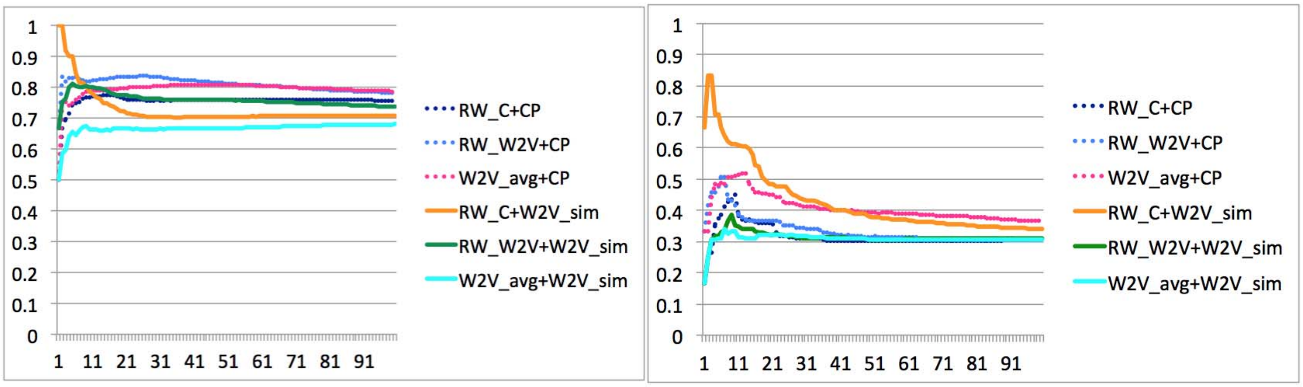


Fig. 7. Left: MAP@1 @100 evaluated by non-medical experts. Right: MAP@1 @100 evaluated by medical experts.

TABLE IV
MAP@100 OF THE DISCOVERED SYMPTOMS BY THE DIFFERENT APPROACHES FOR EACH DISEASE.

	approach	lung cancer	nasopharyngeal carcinoma	diabetes	cirrhosis	colorectal cancer	rectal cancer	avg
Non-expert	RW_C+CP	0.957	0.556	0.809	0.675	**0.779**	**0.759**	0.756
	RW_{W2V}+CP	**0.961**	0.610	0.825	**0.787**	0.765	0.737	0.781
	$W2V_{avg}$+CP	**0.961**	**0.675**	0.832	0.759	0.744	0.744	**0.786**
	RW_C+$W2V_{sim}$	0.673	0.568	**0.939**	0.639	0.745	0.677	0.707
	RW_{W2V}+$W2V_{sim}$	0.678	0.647	0.882	0.739	0.762	0.700	0.735
	$W2V_{avg}$+$W2V_{sim}$	0.581	0.525	0.900	0.659	0.728	0.683	0.679
Expert	RW_C+CP	**0.244**	0.327	0.132	0.281	0.400	0.441	0.304
	RW_{W2V}+CP	0.205	0.402	**0.171**	0.250	0.455	0.363	0.307
	$W2V_{avg}$+CP	0.202	**0.493**	0.141	0.374	**0.533**	**0.454**	**0.366**
	RW_C+$W2V_{sim}$	0.217	0.414	0.153	**0.401**	0.445	0.410	0.340
	RW_{W2V}+$W2V_{sim}$	0.150	0.391	0.128	0.351	0.410	0.424	0.309
	$W2V_{avg}$+$W2V_{sim}$	0.154	0.345	0.115	0.376	0.448	0.394	0.305

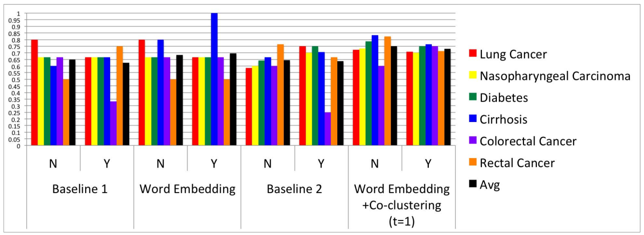


Fig. 8. Precision of conditional terms selection without/with (Y/N) filtering the uncertainty symptoms.

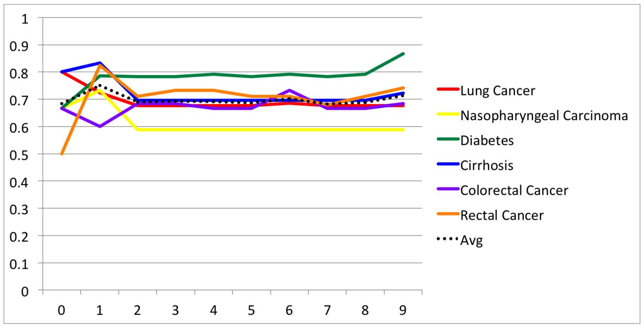


Fig. 9. Precision of conditional terms selection by varying the number of times of co-clustering.

discovered by the word embedding combining co-clustering approach, respectively.

In order to observe the effect of error propagation, we compared the precisions of the conditional terms for all the symptoms discovered in the previous step and the conditional terms for only the symptoms scored by the experts as 1 or 0.5 being remaining. From Fig. 8, it shows that the precisions of the proposed two methods are not affected much no matter the uncertain symptoms are filtered out or not. However, the precisions of the baseline methods do not keep stable for both cases. It also shows that the Word Embedding+Co-clustering approach achieves the highest macro average precision across diseases. About 75% of the discovered conditional terms help further understanding of the relationship between the symptoms and diseases. Therefore, the conditional terms for all the discovered symptoms without filtering are remaining for the following experiments.

Fig. 9 shows the precisions of the discovered conditional terms by varying the number of times of co-clustering from 0 to 9. The results imply that the co-clustering can improve the precisions of the conditional terms for most diseases except `lung cancer` and the `colorectal cancer`. On average, setting the number of times=1 achieves the best performance. According to our observations, the precisions of the result for a disease decreases when performing more times of co-clustering. It is reasonable because when the disease has more candidate conditional terms and context terms, more times of co-clustering will contribute more indirect semantics among the terms. The disease `diabetes` has more focused candidate conditional terms. Accordingly, the precision of the conditional terms for `diabetes` increases up to 0.88 when the number of times of co-clustering is increased to 9. After the 9 times of co-clustering, the disease `diabetes` can find out really helpful conditions such as `skin` and `groin`. For example, “`Diabetic skin pruritus` is

TABLE V
PRECISION OF CONDITIONAL TERMS SELECTION FOR DIFFERENT DISEASES.

approach	lung cancer	nasopharyngeal carcinoma	diabetes	cirrhosis	colorectal cancer	rectal cancer	avg
Baseline 1	**0.800**	0.667	0.667	0.600	**0.667**	0.500	0.650
Word Embedding	**0.800**	0.667	0.667	0.800	**0.667**	0.500	0.684
Baseline 2	0.586	0.600	0.643	0.667	0.600	0.765	0.644
Word Embedding+Co-clustering (t=1)	0.724	**0.733**	**0.786**	**0.833**	0.600	**0.824**	**0.750**
$\#Candidateconditionalterms$	2332	1134	1326	1452	1638	1580	1577

a common clinical complication of diabetes. The clinical manifestation is mainly pruritus." Therefore, diabetes has symptom pruritus on skin, which is a conditional term. Moreover, it is helpful to check the symptom impaired wound healing on the body part of groin when diagnosing the disease diabetes.

In Table V, all types of conditional terms are put together. The word embedding combined with co-clustering approach performs the best, the macro precision cross different diseases is up to 0.75.

VI. CONCLUSION

In this paper, we proposed a system to automatically discover the relationship between diseases and symptoms with conditions from the Internet. The scoring methods are designed to rank the candidate symptoms of a disease extended from the seed symptoms. Embedded representation learning for each candidate conditional term and the co-clustering approach are combined to discover the representative conditional terms. The results of performance evaluation show that the proposed methods can correctly detect the well-known top 6 symptoms of diseases and find the top 100 symptoms with a stable quality 0.78 macro average MAP for the testing diseases. Moreover, the jointly method can effectively discover the conditional terms associated with the symptoms of a disease. In the further, we will extend these strategies to discover the other relationships among the medical concepts with conditional terms.

REFERENCES

[1] L. Sun, C. Liu, C. Guo, H. Xiong, and Y. Xie, "Data-driven automatic treatment regimen development and recommendation," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1865–1874.

[2] M. S. Simpson, E. M. Voorhees, and W. Hersh, "Overview of the trec 2014 clinical decision support track," LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD, Tech. Rep., 2014.

[3] R. Feldman, O. Netzer, A. Peretz, and B. Rosenfeld, "Utilizing text mining on online medical forums to predict label change due to adverse drug reactions," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2015, pp. 1779–1788.

[4] K. Lee, A. Qadir, S. A. Hasan, V. Datla, A. Prakash, J. Liu, and O. Farri, "Adverse drug event detection in tweets with semi-supervised convolutional neural networks," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 705–714.

[5] S. Zhang, T. Kang, L. Qiu, W. Zhang, Y. Yu, and N. Elhadad, "Cataloguing treatments discussed and used in online autism communities," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 123–131.

[6] T. R. Goodwin and S. M. Harabagiu, "Medical question answering for clinical decision support," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 297–306.

[7] D. Savenkov and E. Agichtein, "When a knowledge base is not enough: Question answering over knowledge bases with external text data," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016, pp. 235–244.

[8] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin, "Knowledge base completion via search-based question answering," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 515–526.

[9] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[10] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[11] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, and T. M. Mitchell, "Coupled semi-supervised learning for information extraction," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 101–110.

[12] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell, "Toward an architecture for never-ending language learning." in *AAAI*, vol. 5. Atlanta, 2010, p. 3.

[13] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel *et al.*, "Never-ending learning," *Communications of the ACM*, vol. 61, no. 5, pp. 103–115, 2018.

[14] T. Ruan, M. Wang, J. Sun, T. Wang, L. Zeng, Y. Yin, and J. Gao, "An automatic approach for constructing a knowledge base of symptoms in chinese," *Journal of biomedical semantics*, vol. 8, no. 1, p. 33, 2017.

[15] D. Zhang, D. He, S. Zhao, and L. Li, "Query expansion with automatically predicted diagnosis: iris at trec cds track 2016." in *TREC*, 2016.

[16] Z. Wang, Z. Wang, J. Li, and J. Z. Pan, "Building a large scale knowledge base from chinese wiki encyclopedia," in *Joint International Semantic Technology Conference*. Springer, 2011, pp. 80–95.

[17] M. Li, Y. Shi, Z. Wang, and Y. Liu, "Building a large-scale cross-lingual knowledge base from heterogeneous online wikis," in *Natural Language Processing and Chinese Computing*. Springer, 2015, pp. 413–420.

[18] P. Wang, L. Ji, J. Yan, L. Jin, and W.-Y. Ma, "Learning to extract conditional knowledge for question answering using dialogue," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 277–286.

[19] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 653–658.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] G. Bisson and F. Hussain, "Chi-sim: A new similarity measure for the co-clustering task," in *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*. IEEE, 2008, pp. 211–217.