Journal of Biomedical Informatics 58 (2015) S158–S163

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

An automatic system to identify heart disease risk factors in clinical texts over time

Qingcai Chen [a], Haodi Li [a], Buzhou Tang [a,*], Xiaolong Wang [a], Xin Liu [a], Zengjian Liu [a], Shu Liu [a], Weida Wang [a], Qiwen Deng [b], Suisong Zhu [b], Yangxin Chen [c], Jingfeng Wang [c]

[a] Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China
[b] The Sixth People's Hospital of Shenzhen, Shenzhen 518052, China
[c] Department of Cardiology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou 510120, China

ARTICLE INFO

Article history:
Received 30 January 2015
Revised 22 August 2015
Accepted 1 September 2015
Available online 8 September 2015

Keywords:
Risk factor identification
Clinical information extraction
Heart disease
Machine learning

ABSTRACT

Despite recent progress in prediction and prevention, heart disease remains a leading cause of death. One preliminary step in heart disease prediction and prevention is risk factor identification. Many studies have been proposed to identify risk factors associated with heart disease; however, none have attempted to identify all risk factors. In 2014, the National Center of Informatics for Integrating Biology and Beside (i2b2) issued a clinical natural language processing (NLP) challenge that involved a track (track 2) for identifying heart disease risk factors in clinical texts over time. This track aimed to identify medically relevant information related to heart disease risk and track the progression over sets of longitudinal patient medical records. Identification of tags and attributes associated with disease presence and progression, risk factors, and medications in patient medical history were required. Our participation led to development of a hybrid pipeline system based on both machine learning-based and rule-based approaches. Evaluation using the challenge corpus revealed that our system achieved an *F*1-score of 92.68%, making it the top-ranked system (without additional annotations) of the 2014 i2b2 clinical NLP challenge.

1. Introduction

Heart disease attracts much attention, given its history as the number one cause of death in both women and men throughout the world [1]. Several factors have been identified as risks related to heart disease, including hyperlipidemia, hypertension, obesity, and smoking status. In order to predict and prevent heart disease, it is necessary to first identify risk factors embedded in unstructured clinical documents. Over the last decade, many studies have been undertaken to identify these risk factors, resulting in the creation of publicly available tools, such as clinical Text Analysis and Knowledge Extraction System [2], an open-source tool capable of identifying smoking status. However, no study has investigated the identification of all risk factors associated with heart disease, possibly due to the diversity of their clinical descriptions.

Heart disease is often related to other diseases, such as diabetes, that share several observable characteristics, including obesity and smoking status, as well as some medications, such as metoprolol. All of these were regarded as heart disease risk factors for this study.

The main challenge in identifying all heart disease risk factors is that they are presented in a variety of forms in clinical texts. To comprehensively investigate the identification of all heart disease risk factors, the National Center of Informatics for Integrating Biology and Beside (i2b2) issued a risk factor identification track (track 2) in the clinical natural language processing (NLP) challenge in 2014 [3]. The goal was to identify information medically related to heart disease risk and track its progression over sets of longitudinal patient medical records. We participated in this track and developed a hybrid pipeline system based on both machine learning and rule-based approaches.

In our system, all heart disease risk factors were divided into three categories according to their descriptions, with each category identified individually. Evaluation using the challenge corpus revealed that our system achieved an *F*1-score of 92.86%, making it the top-ranked system (without additional annotations) for the 2014 i2b2 clinical NLP challenge.

* Corresponding author at: Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China.

E-mail addresses: qingcai.chen@gmail.com (Q. Chen), Haodili.hit@gmail.com (H. Li), tangbuzhou@gmail.com (B. Tang), wangxl@insun.hit.edu.cn (X. Wang), hit.liuxin@gmail.com (X. Liu), liuzengjian.hit@gmail.com (Z. Liu), liushuhit@outlook.com (S. Liu), weida.wong@gmail.com (W. Wang), qiwendeng@hotmail.com (Q. Deng), 13809883596@163.com (S. Zhu), tjcyx1995@163.com (Y. Chen), dr_wjf@hotmail.com (J. Wang).

http://dx.doi.org/10.1016/j.jbi.2015.09.002
1532-0464/

2. Related work

The heart disease risk factor identification track of the 2014 i2b2 clinical NLP challenge consisted of two subtasks: risk factor extraction and time attribute identification. To the best of our knowledge, no study has ever been specifically designed for heart disease risk factor identification, although many related studies have been proposed. The most closely related study by Roy et al. developed a hybrid NLP pipeline system to extract Framingham heart failure criteria with time attributes from electronic health records [4].

Heart disease risk factor extraction is a typical information extraction task related to clinical concept recognition [5–9], phenotyping [10], smoking status identification [11–15], obesity identification [16,17], etc. clinical concept recognition is a named entity recognition (NER) task that extracts all problems, treatments, and tests, where problems include diseases and observable characteristics and treatments include medications. The most representative work concerning clinical concept recognition is the 2010 i2b2 clinical NLP challenge, where various machine learning-based, rule-based, and hybrid methods were proposed [18–20]. Phenotypes that include diseases and some observable characteristics have also been widely investigated. Chaitanya et al. summarized approaches for phenotyping [10]. The i2b2 clinical NLP challenges in 2006 and 2008 involved a track on smoking status identification and a track on obesity identification, respectively. The best system for smoking status identification used a method based on support vector machines (SVMs) [21], whereas the best system for obesity identification combined dictionary lookup, rule-based, and machine learning-based methods [17].

The time attribute of each heart disease risk factor represents the relationship between risk factor and the corresponding document creation time (DCT), which is similar to the temporal relationship between a clinical event and DCT in the 2012 i2b2 clinical NLP challenge [22], except that the value of the time attribute can be any combination of {"before", "during", or "after"} rather than a single variable consisting of {"before", "during", "after"}. Most state-of-the-art systems presented for the 2012 i2b2 clinical NLP challenge used machine learning-based methods to extract relationships between events and DCT [23,24]. For example, the best system proposed by Tang et al. adopted SVMs [23].

3. Material and methods

3.1. Dataset

The i2b2 challenge organizers manually annotated longitudinal records of 300 patients (1304 documents), from which 180 patients (790 documents) were used as a training set and the remaining 120 patients (514 documents) were used as a test set. The annotation guidelines defined a set of tags to indicate the presence and progression of diseases (diabetes, heart disease), associated risk factors (hyperlipidemia, hypertension, obesity status, family history, and smoking status), and associated medications. Each tag for the diseases and associated risk factors had one indicator value from its own set, while each tag for the associated medications could have two indicators (denoted by "type1" and "type2") to identify its category.

A brief description of each tag type in the challenge data is presented in Table 1. For more information, please refer to the annotation guidelines [25]. The challenge organizers released the data in two versions: complete and gold. The former provides the evidence (if any exists) for each tag and is used for system development. The latter only provides each tag itself without any evidence and is used for system evaluation.

Fig. 1 shows an example of a tag extracted from sample data (321-03.xml) in both versions, where the evidence associated with the tag is listed in the "text" field.

3.2. Overview of system

Our system identified each type of tag in the following order:

1. Extract evidence (if any exists) by type and indicator.
2. Determine attribute (i.e., time, if it exists).

By analyzing the evidence of tags, we found that the tags mainly fell into the following three categories:

1. Phrase-based tags, the evidence for which is provided explicitly in phrases.
2. Logic-based tags, the evidence for which is provided explicitly in phrases/sentences, but needs additional logical inferences, such as numerical comparisons.
3. Discourse-based tags, the evidence for which is not provided explicitly, but is embedded in clinical text fragments.

Table 2 lists these three categories of evidence-based tags, where the evidence is marked in bold and italics, followed by their tags in parenthesis. The evidence associated with phrase- and logic-based tags are very similar, with the difference being whether logical inference is further required after the phrases are located. For example, the blood pressure (BP) measurement "BP 140/80" is evidence of hypertension due to a high systolic pressure of 140 (Fig. 1). If the BP measurement of a patient is 120/80, "BP 120/80" will not qualify as evidence. Each tag listed in Table 1 may belong to multiple categories mentioned above based on the associated evidence and distinguished by its indicator(s).

The relationships between tag types listed in Table 1 and tag categories listed in Table 2 are shown in Table 3, where each item indicates to which category a tag with an indicator belongs.

Table 1
A brief description of each tag type used in the 2014 i2b2 clinical NLP challenge data.

Tag	Indicator	Attribute	Number	
			Training	Test
Diabetes	Mention, high A1c, high glucose	Time	1695	1180
CAD	Mention, event, test result, symptom	Time	1186	784
Hyperlipidemia	Mention, high cholesterol, high LDL	Time	1062	751
Hypertension	Mention, high blood pressure (high bp)	Time	1926	1293
Obesity status	Mention, BMI,[a] waist circumference	Time	433	262
Family history	Present, no present	NA[b]	790	514
Smoking status	Current, past, ever, never, unknown	NA	771	512
Medication	Metoprolol, . . . , lorquess	Time	8638	5674

[a] Body mass index.
[b] Not available.

Complete Version:
<HYPERTENSION id="DOC9" time="during DCT" indicator="high bp">
<HYPERTENSION id="H6" start="721" end="728" text="150/70." time="during DCT" indicator="high bp"/>
<HYPERTENSION id="H7" start="2998" end="3007" text="BP 140/80" time="during DCT" indicator="high bp"/>
<HYPERTENSION id="H9" start="2998" end="3007" text="BP 140/80" time="during DCT" indicator="high bp"/>
<HYPERTENSION id="H10" start="721" end="727" text="150/70" time="during DCT" indicator="high bp"/>
</HYPERTENSION>
Gold Version:
<HYPERTENSION id="DOC9" time="during DCT" indicator="high bp"/>

Fig. 1. An example of a tag found in both complete and gold versions.

Table 2
Three categories of evidence-based tags.

Category	Evidence
Phrase-based	Continue ***beta blocker*** (Medication), ***CCB*** (Medication)
Logic-based	***BP 140/80*** (Hypertension, high bp), P 72, Wt 276 lb
Discourse-based	***Catheterization revealed his LAD stent was patent but he had a 90% lesion proximal to the stent*** (CAD, event)

Table 3
Relationships between the types and categories of tags in Tables 1 and 2, respectively.

Tag	Phrase-based	Logic-based	Discourse-based
Diabetes	Mention	High A1c, high glucose	NA
CAD	Mention	NA	Event, test result, symptom
Hyperlipidemia	Mention	High cholesterol, high LDL	NA
Hypertension	Mention	High blood pressure	NA
Obesity status	Mention	BMI, waist circumference	NA
Family history	NA	Present, not present	NA
Smoking status	NA	NA	Current, past, . . ., unknown
Medication	Metoprolol, . . ., lorqess	NA	NA

For example, a case of hypertension with a "mention" constitutes a phrase-based tag, while a case of hypertension associated with another indicator constitutes a logic-based tag, as observed in the example from Fig. 1. The proportions of the phrase-,logic-, and discourse-based tags in the training set are 85.33%, 8.10%, and 6.57%, respectively.

After all tags were mapped into the three categories in Table 2, we proposed a unified framework for each category. Fig. 2 shows an overview of our system consisting of six components: a preprocessing module, three tag extraction modules, a time attribute identification module, and a post-processing module.

Given a raw data file containing clinical text, the preprocessing module first performed sentence boundary detection and tokenization. Then the three tag extraction modules extracted evidence of tags from the three categories in Table 2 and determined their type and indicator. Subsequently, the time attribute identification module determined the time attribute of each piece of evidence (if any exists). Finally, the post-processing module converted the tags from the complete version into those from the gold version for evaluation. We used the tokenization module from MedEx [26],[1] a specific tool for medical information extraction, for sentence boundary detection and tokenization. The tag extraction modules and the time attribute identification module are described in detail in the following sections.

[1] https://code.google.com/p/medex-uima/downloads/list.

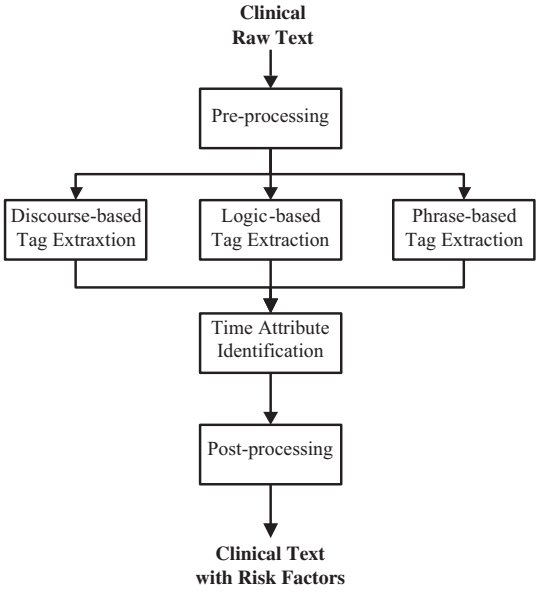

Fig. 2. Overview of our heart disease risk factor identification system.

3.3. Phrase-based tag extraction

Extracting evidence of the phrase-based tags was treated as a NER task in our system. Each piece of evidence was represented by BIOES tags, where B, I, O, and E denote that a token exists at the beginning, middle, outside, or end of a piece of evidence, respectively, and S denotes that the token itself forms a piece of evidence. As an example of evidence from the phrase-based tag in Table 2, the sentence "Continue ***beta blocker, CCB***" was labeled as "Continue/O; beta/B-medication_beta + blockers; blocker/E-medication_beta + blockers; ,/O; CCB/S-medication_calcium-channel + blockers", where "medication" is a type of tag and {"beta blockers", "calcium-channel blockers"} are two indicators of this type of tag. It should be noted that for those medications with two indicators, we only considered the first one (type1).

In this study, we developed an ensemble system for phrase-based tag extraction. The system first used conditional random fields (CRFs) [27] and structured SVMs (SSVMs) [28] to extract evidence from clinical notes individually and then returned the union of their results directly without any treatment of overlapping evidence. The features used in these two base classifiers included bag-of-words, part-of-speech (POS) tags, combinations of tokens and POS tags, sentence information, affixes, orthographical features, word shapes, section information, general NER information, word representation features, dictionary features, and negation information. Except for the dictionary features and negation information, all remaining features were similar to those used in our de-identification system for that challenge [29]. Here, we only introduced word shapes, section information, dictionary features, and

negation information. For other features, please refer to our paper for de-identification [29].

1. Word shapes. We used two typical types of word shapes: one is generated by mapping any uppercase character, lowercase character, digit, or other character in the word to 'A', 'a', '#', or '–', respectively, while the other is generated by mapping consecutive uppercase characters, lowercase characters, digits, or other characters to 'A', 'a', '#', or '–', respectively. For instance, the two types of word shapes for "PO/5mg" are "AA-#aa" and "A-#a".
2. Section information. We extracted 29 section headers, such as "family history", from the training set and determined to which section the word belonged.
3. Dictionary features. We extracted a heart disease-related drug dictionary from DrugBank [30],[2] found all drugs mentioned in the clinical text using dictionary lookup, and checked each token against any drug that we found. For each DrugBank record, we first checked whether it was heart disease-related according to the "name", "synonyms", "brands", and "description" fields. If any medication indicator was mentioned in the "description" field or any medication evidence appeared in the other three fields, this record was regarded as a heart-disease-related record. We then extracted the "name", "synonyms", and "brands" fields of every heart-disease-related record as drug names and the corresponding indictor as the type associated with the drug names. No other dictionary was used in our system outside of this heart-disease-related drug dictionary.
4. Negation information. In order to determine the presence of any negative words in a given sentence, we used the negative word list from NegEx [31].[3]

3.4. Logic-based tag extraction

We first defined two criteria for extracting evidence of logic-based tags and then filtered them by polarity and sentence type. The criteria used for our system included:

1. Numeric constraints: Find all numerical evidence, such as "LDL measurement of over 100 mg/dL", which constitutes evidence of hyperlipidemia with high LDL as determined by "LDL > 100 mg/dL". Each category of logic-based tags contains numerical evidence similar to this.
2. Co-occurrence constraints: Find all evidence determined by multiple keywords, such as "Early-onset CAD in mother", which constitutes evidence of family history as determined by "early, CAD, mother". Only evidence of family history tags were extracted using this criterion in our system.

For the evidence extracted, we removed those in negative or subjunctive mood. The negation information of evidence was determined using NegEx [31] and the subjunctive mood was determined using manually defined rules.

3.5. Discourse-based tag extraction

Given that evidence from discourse-based tags is not explicit, it is difficult to extract them directly, unlike the other two tag categories mentioned above. In this study, we first generated evidence candidate sentences with discourse-based tags according to indicator-related words/phrases, such as symptom-related phrases like "unstable angina", and then checked their indicator-related status using SVMs. Fig. 3 illustrates the extraction process.

The features used in the classifier included term frequency–inverse document frequency (TF–IDF) of words, unigrams of words, bigrams of words, negation information of sentences mentioned in the phrase-based tag extraction module, and negation information of indicator-related words/phrases determined by NegEx. It should be noted that the negation information of words/phrases was used in the logic- and discourse-based tag extraction subsystems, while the negation information of sentences was used in the phrase-based subsystem. This is because target words/phrases have been extracted in advance in the logic- and discourse-based tag extraction subsystems, but they have not in the phrase-based subsystem.

3.6. Time attribute identification

Time attribute identification is actually temporal-relationship extraction for evidence and DCT pairs, which can be recognized as a classification problem similar to the previous study for the 2012 i2b2 challenge [22]. As the temporal relationship between a piece of evidence and DCT in this challenge can be any combination of {"before", "during", or "after"}, the time attribute identification problem is a multi-label classification problem. In this study, we used the label-powerset strategy [32] to convert the multi-label classification problem into a single-label classification problem and applied SVMs toward solving it. The features of a piece of evidence used for time attribute identification included TF–IDF of words, unigrams of words, bigrams of words, evidence tag type, evidence indicator, and temporal relationship between the evidence-related time and DCT. The evidence-related time was the temporal expression nearest to the evidence, such as "yesterday" and "2061/08/20" (a fictional date used for de-identification) in our system. The evidence-related times and DCT were extracted using customized NorMA [33],[4] which is a rule-based temporal-expression normaliser for clinical texts.

4. Results

We used SVM^{hmm},[5] libshortText[6] and CRFsuite [34] as the implementations of SSVMs, SVMs, and CRFs, respectively, and optimized parameters of all classifiers using 10-fold cross-validation on the training set.

Our system was evaluated using the evaluation script provided by the challenge organizers that outputs macro-/micro-precision, -recall, and -*F*1-score, of which micro-precision and -*F*1-score were used as the primary measurements. For this task, a participating team could submit three runs. The results of our best run are reported in this section and shown in Table 4. The best micro-precision, -recall, and -*F*1-scores were 91.06%, 94.36%, and 92.68%, respectively, while the micro-precision, -recall, and -*F*1-scores following 10-fold cross-validation of the training set were 89.92%, 92.20%, and 91.05%, respectively. The micro-*F*1-score of our best run was slightly lower than the best results reported from the challenge (92.68% vs. 92.76%, respectively). Among all tag types, our system did not perform very well for coronary artery disease (CAD), obesity status, and smoking status, with micro-*F*1-scores of 82.53%, 88.57%, and 88.61%, respectively.

For heart disease risk factor extraction, there were three subsystems: phrase-based, logic-based, and discourse-based, corresponding to the three tag categories from Table 2. To test the contribution

[2] http://www.drugbank.ca/.
[3] https://code.google.com/p/negex.
[4] http://www.cs.man.ac.uk/~filannim/.
[5] http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html.
[6] http://www.csie.ntu.edu.tw/~cjlin/libshorttext/.

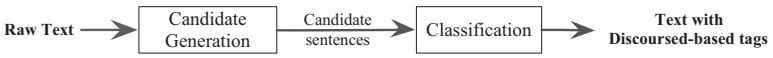

Fig. 3. Discourse-based tag extraction process.

Table 4
Risk factor identification results.

Category	Macro			Micro		
	P	*R*	*F*1	*P*	*R*	*F*1
Diabetes	0.6497	0.6653	0.6574	0.9146	0.9441	0.9291
CAD	0.3556	0.3325	0.3437	0.8440	0.8074	0.8253
Hyperlipidemia	0.4881	0.4460	0.4470	0.9681	0.9308	0.9491
Hypertension	0.7451	0.7459	0.7455	0.9629	0.9373	0.9715
Obesity status	0.1479	0.1540	0.1409	0.8520	0.9008	0.8757
Family history	0.9679	0.9728	0.9703	0.9634	0.9728	0.9681
Smoking status	0.8855	0.9086	0.8969	0.8616	0.9121	0.8861
Medication	0.7712	0.8261	0.7977	0.9012	0.9593	0.9293
All	0.9119	0.9399	0.9257	0.9106	0.9436	0.9268

Table 5
Contribution of each subsystem to global performance.

System	Macro			Micro		
	P	*R*	*F*1	*P*	*R*	*F*1
Pb_{crf}	0.8827	0.7550	0.8138	0.9242	0.817	0.8673
Pb_{ssvm}	0.8924	0.7745	0.8292	0.9200	0.8267	0.8708
Pb_{all}	0.8872	0.7920	0.8369	0.9152	0.8451	0.8788
Pb_{all} + Lb	0.9185	0.8566	0.8865	0.9178	0.8881	0.9027
Pb_{all} + Lb + Db	0.9119	0.9399	0.9257	0.9106	0.9436	0.9268

Table 6
Subsystem performance.

System	Macro			Micro		
	P	*R*	*F*1	*P*	*R*	*F*1
Pb_{all}	0.8838	0.9203	0.9017	0.9257	0.9634	0.9441
Lb	0.8829	0.9147	0.8986	0.8350	0.8785	0.8562
Db	0.8601	0.8419	0.8509	0.8098	0.7808	0.7950

of each subsystem, we started with the phrased-based subsystem as a baseline and then added the other two subsystems gradually. The results are shown in Table 5. The SSVM phrase-based subsystem (Pb_{ssvm}) performed slightly better relative to the CRF phrase-based performance (Pb_{crf}) because of higher recall. The performance of combinations of CRF and SSVM phrase-based systems (Pb_{all}) was superior to any of them individually. When the logic-based subsystem (Lb) was added, the micro-*F*1-score improved by 2.4%. By adding the discourse-based subsystem (Db), the micro-*F*-score improved by an additional 2.4%.

To further investigate the performance of our system on each tag category, we compared the performance of each subsystem individually without considering time attribution (Table 6). Among the three subsystems, the phrase-based subsystem (Pb_{all}) achieved the highest micro-*F*1-score (0.9441), higher than both logic- (Lb) and discourse-based subsystems (Db) by 8.79% and 14.95%, respectively, indicating that our system performed best on phrase-based tags relative to logic- or discourse-based tags.

To test the impact of the time attribute identification module, we compared our system to an upper-boundary system that used the same risk factor extraction module as ours and evaluated the return of correct time attributions (based on the gold standard). The micro-precision, -recall, and -*F*1-scores of the upper-boundary system were 0.9213, 0.9355, and 0.9283, respectively. The *F*1-score difference between the two systems was 0.15%, indicating that the impact of time attribute identification on the overall performance of our system was minimal.

5. Discussion

We developed a pipeline system to identify heart disease risk factors from clinical texts over the course of time and participated in Track 2 of the 2014 i2b2 clinical NLP challenge. The task of this track was to identify diseases, risk factors, medications, and time attributes associated with their presentation relative to DCT. Our system first extracted these risk factors, then identified their time attributes. Based on the evidence characteristics associated with each risk factor, we divided the risk factors into three categories and proposed three individual subsystems for each.

Evaluation on the independent test set provided by the challenge organizers revealed that our system achieved promising results with the highest *F*1-score of 92.68%, which was 0.06% lower than the highest *F*1-score submitted to the challenge. It should be noted that our system performed comparable to that of the highest-performing system while using fewer annotations.

Although the overall performance was promising, our system did not perform well on several tag types, including CAD, obesity status, and smoking status. For the two tag types that contain a large number of discourse-based tags (CAD and smoking status), the proportion of negative samples (not indicator-related) in evidence candidates of discourse-based tags were very high. For example, the proportion of negative samples related to events was 68.4%. The high proportion of negative samples likely explains the relatively poor results. The obesity status tags account for only 2.4% of the tags in the data sets. Their low frequency makes them difficult to identify because of the class imbalance problem associated with machine learning methods. According to the overall report from the i2b2 2014 challenge, other participating systems experienced similar results related to these three tag types.

Possible reasons why the phrase-based tag extraction subsystem outperformed the logic-based tag extraction subsystem are as follows: we did not define accurate rules according to the evidence present in the training set individually and the rules developed from the training set did not completely cover the evidence present in the test set.

For further improvement, there are three possible directions. First, we can refine rules for logic-based tags. Second, given that the logic-based tag extraction subsystem is completely dependent upon rules, we can try machine-learning-based methods. Third, it is worth trying to integrate medical domain knowledge, such as Unified Medical Language System [35] and Systematized Nomenclature of Medicine-Clinical Terms [36], to avoid generating too many negative samples in the discourse-based tag extraction subsystem by limiting the candidates to those that contain certain concepts of medical domain knowledge.

6. Conclusion

In this study, we developed a hybrid pipeline heart disease risk factor identification system for clinical texts that can identify diseases, associated risk factors, associated medications, and the time they are presented. In our system, all heart disease risk factors were divided into three categories according to their descriptions, with each category identified individually. With this system, we participated in Track 2 of the 2014 i2b2 clinical NLP challenge. Our method achieved a micro-*F*1-score of 92.68%, which was competitive with other state-of-the-art systems.

Conflict of interest

None declared.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This paper is supported in part by grants: NSFCs (National Natural Science Foundations of China) (Grant Nos.: 61402128, 61473101, 61173075, and 61272383) and Strategic Emerging Industry Development Special Funds of Shenzhen (Grant Nos.: JCYJ20140508161040764, JCYJ20140627163809422 and JCYJ201417172417105). We also thank the 2014 i2b2 NLP challenge organizers for making the annotated data set available.

References

[1] A.S. Go, D. Mozaffarian, V.L. Roger, E.J. Benjamin, J.D. Berry, M.J. Blaha, S. Dai, E. S. Ford, C.S. Fox, S. Franco, H.J. Fullerton, C. Gillespie, S.M. Hailpern, J.A. Heit, V.J. Howard, M.D. Huffman, S.E. Judd, B.M. Kissela, S.J. Kittner, D.T. Lackland, J.H. Lichtman, L.D. Lisabeth, R.H. Mackey, D.J. Magid, G.M. Marcus, A. Marelli, D.B. Matchar, D.K. McGuire, E.R. Mohler, C.S. Moy, M.E. Mussolino, R.W. Neumar, G. Nichol, D.K. Pandey, N.P. Paynter, M.J. Reeves, P.D. Sorlie, J. Stein, A. Towfighi, T. N. Turan, S.S. Virani, N.D. Wong, D. Woo, M.B. Turner, American heart association statistics committee and stroke statistics subcommittee, heart disease and stroke statistics–2014 update: a report from the American Heart Association, Circulation 129 (3) (2014) e28–e292, http://dx.doi.org/10.1161/01.cir.0000441139.02102.80.

[2] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inf. Assoc. 17 (5) (2010) 507–513.

[3] S. Amber, K. Christopher, X. Hua, Ö. Uzuner, Practical applications for NLP in Clinical Research: the 2014 i2b2/UTHealth shared tasks, J. Biomed. Inform. 58S (2015) S1–S5.

[4] R.J. Byrd, S.R. Steinhubl, J. Sun, S. Ebadollahi, W.F. Stewart, Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records, Int. J. Med. Inf. <http://www.sciencedirect.com/science/article/pii/S1386505612002468>.

[5] Ã. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inf. Assoc.: JAMIA 18 (5) (2011) 552–556, http://dx.doi.org/10.1136/amiajnl-2011-000203.

[6] M. Torii, K. Wagholikar, H. Liu, Using machine learning for concept extraction on clinical documents from multiple data sources, J. Am. Med. Inf. Assoc.: JAMIA 18 (5) (2011) 580–587, http://dx.doi.org/10.1136/amiajnl-2011-000155.

[7] B. Tang, H. Cao, Y. Wu, M. Jiang, H. Xu, Clinical entity recognition using structural support vector machines with rich features, in: Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics, DTMBIO '12, ACM, New York, NY, USA, 2012, pp. 13–20. doi:http://dx.doi.org/10.1145/2390068.2390073, http://doi.acm.org/10.1145/2390068.2390073.

[8] A.N. Nguyen, M.J. Lawley, D.P. Hansen, R.V. Bowman, B.E. Clarke, E.E. Duhig, S. Colquist, Symbolic rule-based classification of lung cancer stages from free-text pathology reports, J. Am. Med. Inf. Assoc.: JAMIA 17 (4) (2010) 440–445, http://dx.doi.org/10.1136/jamia.2010.003707.

[9] L. Cui, A. Bozorgi, S.D. Lhatoo, G.-Q. Zhang, S.S. Sahoo, EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification, in: AMIA Annual Symposium Proceedings/AMIA Symposium, AMIA Symposium 2012, 2012, pp. 1191–1200.

[10] C. Shivade, P. Raghavan, E. Fosler-Lussier, P.J. Embi, N. Elhadad, S.B. Johnson, A. M. Lai, A review of approaches to identifying patient phenotype cohorts using electronic health records, J. Am. Med. Inf. Assoc.: JAMIA 21 (2) (2014) 221–230, http://dx.doi.org/10.1136/amiajnl-2013-001935.

[11] Ö. Uzuner, I. Goldstein, Y. Luo, I. Kohane, Identifying patient smoking status from medical discharge records, J. Am. Med. Inf. Assoc.: JAMIA 15 (1) (2008) 14–24, http://dx.doi.org/10.1197/jamia.M2408.

[12] A.M. Cohen, Five-way smoking status classification using text hot-spot identification and error-correcting output codes, J. Am. Med. Inf. Assoc.: JAMIA 15 (1) (2008) 32–35, http://dx.doi.org/10.1197/jamia.M2434. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274879/>.

[13] G.K. Savova, P.V. Ogren, P.H. Duffy, J.D. Buntrock, C.G. Chute, Mayo clinic NLP system for patient smoking status identification, J. Am. Med. Inf. Assoc.: JAMIA 15 (1) (2008) 25–28, http://dx.doi.org/10.1197/jamia.M2437. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274870/>.

[14] R. Wicentowski, M.R. Sydes, Using implicit information to identify smoking status in smoke-blind medical discharge summaries, J. Am. Med. Inf. Assoc.: JAMIA 15 (1) (2008) 29–31, http://dx.doi.org/10.1197/jamia.M2440. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274867/>.

[15] D.T. Heinze, M.L. Morsch, B.C. Potter, R.E. Sheffer, Medical i2b2 NLP smoking challenge: the a-life system architecture and methodology, J. Am. Med. Inf. Assoc.: JAMIA 15 (1) (2008) 40–43, http://dx.doi.org/10.1197/jamia.M2438. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274871/>.

[16] Ö. Uzuner, Recognizing obesity and comorbidities in sparse data, J. Am. Med. Inf. Assoc.: JAMIA 16 (4) (2009) 561–570, http://dx.doi.org/10.1197/jamia.M3115.

[17] H. Yang, I. Spasic, J.A. Keane, G. Nenadic, A text mining approach to the prediction of disease status from clinical discharge summaries, J. Am. Med. Inf. Assoc.: JAMIA 16 (4) (2009) 596–600, http://dx.doi.org/10.1197/jamia.M3096.

[18] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, X. Zhu, Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010, J. Am. Med. Inf. Assoc.: JAMIA 18 (5) (2011) 557–562, http://dx.doi.org/10.1136/amiajnl-2011-000150.

[19] M. Jiang, Y. Chen, M. Liu, S.T. Rosenbloom, S. Mani, J.C. Denny, H. Xu, A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, J. Am. Med. Inf. Assoc.: JAMIA 18 (5) (2011) 601–606, http://dx.doi.org/10.1136/amiajnl-2011-000163.

[20] S.R. Jonnalagadda, D. Li, S. Sohn, S.T.-I. Wu, K. Wagholikar, M. Torii, H. Liu, Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules, J. Am. Med. Inf. Assoc.: JAMIA 19 (5) (2012) 867–874, http://dx.doi.org/10.1136/amiajnl-2011-000766. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3422831/>.

[21] C. Clark, K. Good, L. Jezierny, M. Macpherson, B. Wilson, U. Chajewska, Identifying smokers with a medical extraction system, J. Am. Med. Inf. Assoc.: JAMIA 15 (1) (2008) 36–39, http://dx.doi.org/10.1197/jamia.M2442. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274874/>.

[22] W. Sun, A. Rumshisky, Ö. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 challenge, J. Am. Med. Inf. Assoc. (2013) (amiajnl–2013).

[23] B. Tang, Y. Wu, M. Jiang, Y. Chen, J.C. Denny, H. Xu, A hybrid system for temporal information extraction from clinical text, J. Am. Med. Inf. Assoc.: JAMIA 20 (5) (2013) 828–835, http://dx.doi.org/10.1136/amiajnl-2013-001635.

[24] J. D'Souza, V. Ng, Classifying temporal relations in clinical data: a hybrid, knowledge-rich approach, J. Biomed. Inf. 46 (2013) S29–S39, http://dx.doi.org/10.1016/j.jbi.2013.08.003. <http://www.sciencedirect.com/science?_ob=DownloadURL&_method=finish&_eidkey=1-s2.0-S1532046413001184&count=1&_docType=FLA&md5=57e8062226ffb1890533a158219b623e>.

[25] A. Stubbs, Ö. Uzuner, Annotating risk factors for heart disease in clinical narratives for diabetic patients, J. Biomed. Inform. 58S (2015) S78–S91. doi: http://dx.doi.org/10.1016/j.jbi.2015.05.009.

[26] H. Xu, S.P. Stenner, S. Doan, K.B. Johnson, L.R. Waitman, J.C. Denny, MedEx: a medication information extraction system for clinical narratives, J. Am. Med. Inf. Assoc.: JAMIA 17 (1) (2010) 19–24, http://dx.doi.org/10.1197/jamia.M3378.

[27] J. Lafferty, A. McCallum, F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Departmental Papers (CIS) <http://repository.upenn.edu/cis_papers/159>.

[28] T. Joachims, T. Finley, C.-N.J. Yu, Cutting-plane training of structural SVMs, Mach. Learn. 77 (1) (2009) 27–59, http://dx.doi.org/10.1007/s10994-009-5108-8. <http://dx.doi.org/10.1007/s10994-009-5108-8>.

[29] Z. Liu, Y. Chen, B. Tang, X. Wang, Q. Chen, H. Li, J. Wang, Q. Deng, S. Zhu, Automatic de-identification of electronic medical records using token-level and character-level conditional random fields, J. Biomed. Inform. 58S (2015) S47–S52. doi:http://dx.doi.org/10.1016/j.jbi.2015.06.009.

[30] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, DrugBank 4.0: shedding new light on drug metabolism, Nucl. Acids Res. 42 (D1) (2014) D1091–D1097. <http://nar.oxfordjournals.org/content/42/D1/D1091.short>.

[31] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inf. 34 (5) (2001) 301–310, http://dx.doi.org/10.1006/jbin.2001.1029. <http://www.sciencedirect.com/science/article/pii/S1532046401910299>.

[32] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837, http://dx.doi.org/10.1109/TKDE.2013.39.

[33] A. Kovacevic, A. Dehghan, M. Filannino, J.A. Keane, G. Nenadic, Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives, J. Am. Med. Inf. Assoc. 20 (5) (2013) 859–866.

[34] N. Okazaki, CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007 <http://www.chokkan.org/software/crfsuite/>.

[35] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucl. Acids Res. 32 (suppl 1) (2004) D267–D270. <http://nar.oxfordjournals.org/content/32/suppl_1/D267.short>.

[36] M.Q. Stearns, C. Price, K.A. Spackman, A.Y. Wang, SNOMED clinical terms: overview of the development process and project status, in: Proceedings of the AMIA Symposium, 2001, pp. 662–666 <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243297/>.