Journal of Biomedical Informatics 58 (2015) S183–S188

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Using local lexicalized rules to identify heart disease risk factors in clinical notes

George Karystianis [a,e], Azad Dehghan [a,e], Aleksandar Kovacevic [b], John A. Keane [a,d], Goran Nenadic [a,c,d,*]

[a] School of Computer Science, University of Manchester, Manchester, UK
[b] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[c] Health eResearch Centre, The Farr Institute of Health Informatics Research, Manchester, UK
[d] Manchester Institute of Biotechnology, University of Manchester, UK
[e] The Christie NHS Foundation Trust, Manchester, UK

ARTICLE INFO

Article history:
Received 15 February 2015
Revised 10 June 2015
Accepted 22 June 2015
Available online 29 June 2015

Keywords:
Text mining
Risk factors
Heart disease
Vocabularies
Rule-based modelling

ABSTRACT

Heart disease is the leading cause of death globally and a significant part of the human population lives with it. A number of risk factors have been recognized as contributing to the disease, including obesity, coronary artery disease (CAD), hypertension, hyperlipidemia, diabetes, smoking, and family history of premature CAD. This paper describes and evaluates a methodology to extract mentions of such risk factors from diabetic clinical notes, which was a task of the i2b2/UTHealth 2014 Challenge in Natural Language Processing for Clinical Data. The methodology is knowledge-driven and the system implements local lexicalized rules (based on syntactical patterns observed in notes) combined with manually constructed dictionaries that characterize the domain. A part of the task was also to detect the time interval in which the risk factors were present in a patient.

The system was applied to an evaluation set of 514 unseen notes and achieved a micro-average *F*-score of 88% (with 86% precision and 90% recall). While the identification of CAD family history, medication and some of the related disease factors (e.g. hypertension, diabetes, hyperlipidemia) showed quite good results, the identification of CAD-specific indicators proved to be more challenging (*F*-score of 74%). Overall, the results are encouraging and suggested that automated text mining methods can be used to process clinical notes to identify risk factors and monitor progression of heart disease on a large-scale, providing necessary data for clinical and epidemiological studies.

1. Introduction

Heart disease is the leading cause of death globally [1]: in the UK, for example, about one in six men and one in ten women die from heart disease [2]. Furthermore, a significant part of the human population lives with it (e.g., 2.3 million people in the UK alone). Many studies have been conducted to improve treatment and identify possible risk factors and life-style habits that may make a person more likely to develop heart disease. For example, obesity, coronary artery disease (CAD), hypertension, hyperlipidemia, diabetes, smoking, family history of premature CAD, unhealthy diet and age above 55 have been acknowledged as important risk factors [3]. The ability to identify such risk factors for individual patients is important for both disease prevention and treatment; furthermore, extraction of such information on a large-scale (e.g., from electronic health records (EHRs)) is key for epidemiological studies and understanding the development of the disease.

While EHRs contain coded (structured) information that is undoubtedly useful for such studies, clinical narratives (e.g., letters, doctor notes) are in an unstructured, free-text form and often include rich, contextual information that is not present elsewhere. Processing of such information has been a focus of clinical text mining for over 30 years [4–7], with notable results in harvesting important clinical concepts and events. Efforts have focused on the identification of various concepts, combining a variety of approaches. For example, Goryachev et al. [8] recognized family history from discharge summaries and outpatient clinic notes through a rule-based approach with an *F*-score of 95%, while Wang [9] extracted findings and medical procedures in clinical progress notes by applying both a rule-based system and modelling a conditional random field classifier, with *F*-scores of 49% and 82% respectively. Several approaches have been developed

* Corresponding author at: School of Computer Science, Oxford Road, Manchester M13 9PL, UK. Tel.: +44 161 2756289.
E-mail address: g.nenadic@manchester.ac.uk (G. Nenadic).

http://dx.doi.org/10.1016/j.jbi.2015.06.013
1532-0464/

for identification of medication information from clinical notes. Patrick et al. [10], for example, applied a hybrid approach of supervised learning and rules, while Spasic et al. [11] used a rule-based methodology and Yang [12] mainly relied on a dictionary-based method. Other work has focused on the extraction of medical problems, treatments and tests from clinical narratives with relatively good results, typically with an *F*-score of around 80%. For example, Rink et al. [13] and Jonnalagadda et al. [14] applied machine learning, whereas Xu et al. [15] combined machine learning and rules for that task. Finally, there has been previous work on extracting risk factors for certain conditions: for example, Fiszman and colleagues [16] used a semantic processor that recognized predications to extract metabolic syndrome risk factors (such as obesity, high density lipoprotein, elevated blood pressure) from MEDLINE abstracts, with an overall *F*-score of 59%.

Several community challenges in clinical text processing have been organized to assess the state-of-the-art for specific tasks, including, for example, medication identification [17–19], extraction of co-morbidities [20], etc. One of the tasks in the 2014 i2b2/UThealth Challenge in Natural Language Processing for Clinical Data aimed to identify potential risk factors for heart disease from clinical notes of diabetic patients [3]. The task focused on eight classes, including mentions of *CAD* or factors that are associated with its onset (*diabetes*, *obesity*, *hyperlipidemia*, *hypertension*, *smoking status*, family *history* of CAD and related *medications*). In this paper we describe and evaluate our approach to the task, which uses local lexicalized rules combined with manually constructed dictionaries that characterize the domain. We demonstrate that the rule-based approach is feasible and can be used for reliable large-scale data harvesting.

2. Materials and methods

2.1. Task and data

The task focused on document-level extraction of the eight classes listed above, where each class is characterized by attributes (see Table 1). The five disease classes (CAD, diabetes, obesity, hyperlipidemia, hypertension) are recognized through either explicit presence (mention) of the disease or the progression of clinical markers suggesting the targeted disease (e.g., "hemoglobin levels above 6.5" and "glucose levels over 126" are indicators for diabetes). Different diseases have a different number of indicators: for example, CAD has four (a mention of the disease or its symptom – e.g., angina, event – e.g., heart catheterization, or test – e.g., stress test), obesity has three (mention, body mass index (BMI) and waist circumference (WC)), while hypertension has two (mention and high blood pressure).

The medication class has two attributes: type 1 is the drug category to which the medication belongs (a total of 22, e.g., "sulfonylureas", "meglitinides") and type 2 indicates drugs that can be included in more than one category (e.g., "zestoretic" has type 1 of "ACE inhibitor" and type 2 of "diuretic").

The time attribute refers to the temporal interval in which a risk factor was present in the patient's medical history: before the Document Creation Time (DCT, i.e. the time when the clinical note was created), during DCT and/or after DCT. DCT is considered as an attribute in all of the disease factors and in the medication class. We note that a specific risk factor can be present before, during and after DCT, or in any combination of these.

The smoker class has a status attribute that indicates whether the person is a "current", "past", "ever", or "never" smoker, or if their smoking status is "unknown". Finally, the family history contains the "present" or "not present" indicator that specifies whether the patient has first degree relatives (e.g., parents, siblings) who were diagnosed prematurely with CAD.

The overall task was to indicate the presence of these risk factors at the document level. Specifically, for the five disease factors, the task included a binary, document-level classification (present/absent) for each of the associated indicators and also for the explicit disease mentions. The time attribute further specifies the timeframe(s) (before, during, after). Medication information includes the two types and time (3 values), whereas family history of CAD is a binary classification task (present/absent). Finally, the smoking status needs to be instantiated with one of the five possible values. The task organizers provided a training set (790 clinical notes) and 514 notes as an evaluation set, all fully annotated at the document level [3]. The data are available at the following link https://www.i2b2.org/NLP/DataSets.

2.2. Method overview

After an initial analysis of the training set where we observed common lexical patterns that indicate the presence of the targeted factors (e.g., "male with hypertension", "pmh: diabetes, hypertension"), we designed and implemented a lexicalized rule-based approach for their recognition. Our methodology consists of four steps:

Step 1: creation of specific vocabularies for each class.
Step 2: design and implementation of rules to capture risk factors of interest at the mention level.
Step 3: integration of the mention-level results at the document level.
Step 4: designating the time value to the identified factors.

In the first step, a number of task-specific semantic groups have been identified and lexicalized through a set of custom-made vocabularies that were engineered from open clinical resources (see Table 2). The dictionaries were manually tailored by observing the training set for the usage of terms describing the associated risk factors and expressions related to their indicators (e.g., "blood pressure", "high blood pressure", "systolic blood pressure", etc.), and by adding clinical synonyms and acronyms from the Unified Medical Language System [21] (UMLS) for specific terms of interest.

In the second step, these dictionaries are used to anchor and constrain a generic set of local rules for identification of disease and risk factor mentions by using:

- specific semantic groups, recognized by the vocabularies and/or regular expressions,
- semi-frozen lexical expressions (e.g. "*the patient was diagnosed with*") that are used as anchors for specific entity and attribute types.

We note that the vocabularies used are task-specific, whereas the general rule patterns are focused on the identification of disease risk factors, which are then used to "infer" the mention of a specific disease type (e.g., based on the specific indicators). Generally speaking, the rules were based on structured patterns consisting of semi-frozen syntactic chunks (e.g. noun phrases, verbs and prepositions) and/or semantic place holders (through dictionary mentions), either suggesting the presence of a disease or associated event (e.g., "history of present illnesses include *hypertension*", "underwent *catheterization*", "*stress test* was positive") or specific indicators (e.g., via specific measurements implemented through regular expressions (e.g., "BMI: *NUMBER*", "blood pressure: *NUMBER/NUMBER*"). For example, a rule that captures

Table 1
Heart disease risk factors with their attributes as used in the challenge. The table includes examples and the number of mentions in the training set. Indicators are specific markers that indicate the factor's presence in a patient. Time suggests the period in which the risk factor was present with regards to the creation of the clinical note. The bold parts indicate the targeted mentions.

Risk factor	Attributes		Example	Number of mentions
	Indicator	Time		
Hyperlipidemia	Disease mention	Before, during, after DCT	"PMH: S/p mechanical aortic valve replacement, CHF, HTN, **hyperlipidemia**"	340
	High cholesterol		"patient's **Chol 179**"	5
	Low-density lipoprotein		"**LDL 119**"	33
Hypertension	Disease mention	Before, during, after DCT	"Medications include for **hypertension**, diabetes, and hypercholesterolemia"	524
	High blood pressure		"**Blood pressure: 150/92**"	33
Diabetes	Disease mention	Before, during, after DCT	"PAST MEDICAL HISTORY: Remarkable for seizure, **type II diabetes mellitus**, panhypopituitarism secondary"	524
	Glucose levels		"**glu 192**"	24
	Haemoglobin levels		"**Hemoglobin a1c 8.3**"	101
Obesity	Disease mention	Before, during, after DCT	"Diabetes mellitus Hypertension **Obesity**"	147
	bmi		"**BMI 30.3**"	16
	Waist circumference		–	0
CAD	Disease mention	Before, during, after DCT	"In addition, **CAD**, diabetes, hypertension, CHF"	261
	Event		"She was treated for **NSTEMI** with ASA 325 mg"	237
	Test		"and a **stress test** suggesting anterior ischemia"	74
	Symptom		"cardiac catheterization laboratory because of progressive worsening **angina**"	68
Medication	Type 1	Before, during, after DCT	"Medications **Lisinopril**"	3,085
	Type 2		"Medications **Avandamet**"	13
Family history of CAD	Present, not present		"**Mother diagnosed with cad**"	22
Smoker	Status – current		**Currently smoking** a pack per day	57
	Status – past		**Ex-smoker**	149
	Status – never		"**smoking – no**"	184
	Status – ever		He **smoked once**	9
	Status – unknown			373

mentions of CAD-specific surgery would have two parts: a semi-frozen verbal expression (e.g. various forms of "undergo") followed by a mention matched by the surgery dictionary (as mentioned in Table 2). We have also implemented concept enumeration as it appears quite frequently in the training data, particularly for disease and medication mentions (e.g., "pmhx: *dm*, *htn*, *dementia*", "Medications: *Lisinopril Pravachol*"). Table 3 presents examples of rules for some risk factors. The number of rules for specific entity types (see Table 4) roughly indicates the complexity of the targeted information i.e., the number of associated indicator types. For the design and implementation of the rules we used MinorThird [22], an information extraction development environment that we have previously used for clinical text mining [11].

We note that a document in this task was a set of clinical notes for a given patient and that we are interested whether a risk factor is mentioned or not within the document. Therefore, in the third step, we have integrated the data identified at the mention level to the document level. For example, if we have detected any high blood pressure indicators (e.g., "*bp 150/90 mm/hg*" or "*blood pressure: 160/90 mm/hg*") in a note, we consider that the patient has "hypertension", with an indicator of "high blood pressure" tagged at the document level.

This approach was followed for all entity types and attributes apart from the time dimension. As the clinical notes were longitudinal, there was a high chance that patients have a number of diseases (and indicators) before, during and after the DCT. This is also likely to be the case with the (majority of) administered medications. This hypothesis was confirmed by the training set: from 1223 disease mentions (at the document level), only 15 (1.23%) did not have all three time attributes values (i.e. before DCT, during DCT, after DCT); for medication mentions, from a total of 2191, only 203 (9.26%) had either one or two time attributes. Therefore, we decided to set, as a default, all three values for the time attribute for all the disease and medication mentions, and aim to identify only explicit localized expressions (e.g., "stop drug", "start on drug") to alter these if necessary. Specific disease indicators were treated by different defaults. For example, body mass index and high blood pressure are typically recorded during the creation of the narrative and rarely denote past or future values; we therefore decided to assign the default value of "during DCT" to their time attribute. Other tests (e.g. levels of hemoglobin, glucose, high LDL, high cholesterol, CAD test, CAD symptom, CAD event) typically happen before the date of the current note and hence were set to the default value of "before DCT". This was also supported by the data in the training corpus.

3. Results

The system was formally evaluated as part of the i2b2 challenge. Table 5 displays the summarized results across all data sets

Table 2
Dictionaries used for the lexicalization of rules. A total of 21 dictionaries were manually curated.

Dictionary	Example terms	Size
Haemoglobin	hgblc, hemoglobin, glycohemoglobin, hbg	14
Diabetes	Type 2 diabetes, insulin depedent diabetes, non-insulin-depedent diabetes, adult onset diabetes	66
Hyperlipidemia	hld, hypercholesterolemia, hyperlipoproteinemia, dyslipidemia	14
CAD	cad, coronary artery disease, three-vessel coronary artery disease, heart disease	11
Hypertension	htn, essential hypertension, hypertension, hypertensive disorder	9
CAD symptom	Chest pressure, angina, substernal chest pain, intermitten angina, mild chest discomfort	40
Myocardial infarction	Anteroseptal mi, lateral myocardial infarction, prior inferior myocardial infarction	49
Surgery	Angioplasty, coronary artery bypass, cardiac bypass graft surgery, poba, 2v cabg	60
Smoking concepts	Tobacco use, cigarette smoking, tobacco smoking, cigarette abuse, cigar abuse	28
former smoker	Former smoker, ex smoker, prior smoker, remote smoker, former heavy smoker	10
Obesity	Central obesity, adiposity, obese, general obesity, obesity, morbid obesity,	7
Blood pressure	Blood pressure, bp, sbp, dbp, blood pressures, systolic blood pressure, hbp	7
Gender	Lady, gentleman, man, woman, patient, male, female, f, m	9
History	Past medical history, pmh, pmhx, history, background history, previous history	10
Social activity	Alcohol consumption, alcohol use, substances, substance abuse, drinking, narcotics	17
Medication head	Ointment, inhaler, nebulizer, nebs, puffer, sulphate, cream, paste, elixir, lotion	47
CAD relative	Brother, mother, father, sister, children, son, daughter,	7
Catheter	Left heart catherization, cardiac catherization, cardiac cath	10
CAD stent	RCA stent, aterial stent, taxus stent, cardiac stent, right coronary stent, cypher stent	10
CAD test	Stress test, stress mibi, thallium stress, exercise tolerance test, mibi	5
Diseases	Osteoarthritis, depression, Parkinson's disease, glaucoma, attention deficit disorder	100

(training, development and evaluation). The overall micro precision was 85.57% with recall of 90.07% and a micro *F*-score of 87.76%. We note that there was only a marginal drop in the performance compared to the training data, suggesting that the lexicalized rules managed to generalize the risk factor identification quite well. Table 6 shows the results per entity class for the evaluation set. The highest *F*-score was returned for family history (95.91%), with the highest recall of 96.97% for medication. With the exception of CAD which proved to be the most challenging class to recognize (*F*-score of 73.63%), all other classes were identified with an *F*-score above 85% indicating that the approach we followed was effective in the identification of several components of CAD risk factors in clinical narratives.

4. Discussion

The system's micro *F*-score of 87.76% ranked the system 9th out of the 48 submissions (up to 3 submissions per a team). We note that the performance of the rule-based approach was well above the challenge mean (81.5%) and 5% less than the highest ranking system. This suggests that a rule-based approach for the recognition of heart disease risk factors and the assignment of a time indicator regarding their progression (or not) is worthwhile. To perform an analysis of false positives (FPs) and negatives (FNs), we took a random sample of ten documents for each class from the evaluation set and observed the common types of error that the system generated.

4.1. False positives

A quarter (10/48) of FPs originated from disease mentions that are either related to the family of the patient (e.g., "*family history: diabetes, fh: – **dm**: father*"), are negated (e.g., "*no history of **hypertension***") or refer to allergies (e.g. "*Allergies: **sulfa** drugs*"). It is interesting that negation was not that frequent, contributing to only 6% of cases. Another quarter of FPs (11/48) were ambiguous cues: for example, "*lipids*" and "*ht*" are often used to describe the diagnosis and the status of hyperlipidemia and hypertension respectively, but they can also be used in a different context (e.g., "***lipids** will be checked*", "**ht** *1.82 cm*"). "*Insulin*" was another frequent example, as it refers to both disease mentions (e.g., "**insulin** dependent diabetes mellitus") and a medication. We note that over half of FPs (27/48) are risk factors possibly missed in the annotation process (e.g. "*hemoglobin a1c 7.7*" was not annotated as an indicator for CAD; similarly "*abd: **obese**, non-protuberant*", "***hbalc** 09/20/2065 **6.50***", "*glu 200-265*", "***obese** older gentleman*", "*medications: baclofen, atenolol, lactulose and **lasix***").

4.2. False negatives

In a number of cases, the system ignored disease mentions in particular. For example, CAD attributes (*event*, *test*, *mention*, and *symptom*) were much more variable compared to other classes (e.g., hypertension or hyperlipidemia) and a number of mentions were missed as the rules (although the largest in number) were not flexible enough (25 out of 65 cases) or lexical/variation coverage was limited (e.g. unknown abbreviations, 16/65 cases: for example, "*3 vessel coronary artery bypass surgery*" has appeared in a number of variants, including "*3-vessel ca bypass surgery*", "*3-v ca bypass surgery*" or "*3v bypass surgery*"). This suggests that an extension of the vocabularies could lead to an improvement towards the system's performance. Furthermore, some of the rules used enumerations of diseases (e.g., "*medical issues include* list-of-diseases"), but the cases where a mention was not recognized (e.g., "*diverticulosis*", "*seronegative ra*") would trigger a termination of the enumerated list (and thus the end of the rule match) and as a result a number of mentions were missed. Finally, some particular indicators required clinical background knowledge (e.g., "**LDL 111**" as a high LDL indicator for hyperlipidemia), which was not encoded.

4.3. Time attribute

The implementation of the time attribute default values for the medication and disease mentions has also contributed to some FPs and FNs. As expected, due to the application of the default rule (assigning all three time attribute values to disease and medication mentions), we found FP time attribute values for disease mentions only in nine out of 514 notes (1.75%). In addition, for medication mentions, we detected 144 documents containing FPs, resulting in a lower precision for medications (82%) when compared to the other classes. Although we have implemented some exceptions (e.g., "stop drug", "start on drug"), there were cases where these further required handling medication enumerations (e.g., in "*Patient was immediately told to stop both her Roxicet and* **Monopril**" we correctly time-framed *Roxicet* but *Monopril* was on the default rule, making both [Monopril, during DCT] and [Monopril, after DCT] false positives). Still, this approach contributed to the highest recall (97%) for the medication mentions. Overall, the decision to implement default rules for the time

Table 3
Examples of rules for the recognition of heart disease risk factors.

class	examples	identified span			
	abstract rule rule example	past-medical-history (NP) @history	any token any{0,4}	preposition re('(of|for)')	**disease mention (NP)** **[@diabetes]**
diabetes (disease mention)	His past medical history is also positive for **non-insulin-dependent diabetes mellitus**, aortic valve	past medical history	is also positive	for	**non-insulin diabetes mellitus**

	abstract rule example rule	undergo (verb) re('undergone|underwent|undergo')	any token any{0,1}	**CAD related surgery (NP)** **[@surgery]**
CAD (event)	Since I saw Ms. Law, she underwent a **3-vessel coronary artery bypass surgery**	underwent	a	**3-vessel coronary artery bypass surgery**

	abstract rule example rule	blood pressure (NP) [@pressure	punctuation? a(punctuation)?	**numeric regex** **re('[1-3]?[0-9]** **[0-9]')**	**punctuation** **a(punctuation)**	**numeric regex** **re('[1-3]?[0-9]** **[0-9]')]**
Hyper-tension (blood pressure)	Vital signs: **blood pressure 192/94**	**blood pressure**	:	192	/	94

	abstract rule example rule	gender (NP) @gender	preposition re('(with|w)')	past-medical-history (NP) @history?	disease mention (NP) @disease?	**disease mention (NP)** **[@hyperlipidemia]**
Hyper-lipidemia (cholesterol)	He is a 61-year-old man with cad, dm, **high cholesterol**, htn and family history of early cad	man	with		cad, dm,	**high cholesterol**

Examples show both an "abstract" description of the rule and the MinorThird notation. Rule components in square brackets are the extracted (target) spans that denote the mention of interest; the rest of the rule (if any) specifies the context. The rules use explicit matching of spans (e.g., eq('past') matches string 'past'), regular expressions (re) for matching specific frozen expressions and clues and the vocabularies that contain mentions of specific dictionaries. For example, @surgery includes various surgical procedures, @pressure has variations of blood pressure and @history contains expressions that suggest a mention of history of disease (see Table 2). "Any" matches a given number of tokens (e.g. any{0,4} matches up to 4 tokens).

Table 4
The number of rules created for each of the targeted risk factors.

Risk factors	Number of rules
Medication	10
Hyperlipidemia	66
Hypertension	70
Diabetes	91
Obese	63
CAD	133
Family history of CAD	21
Smoker	94

Table 5
Results per data sets. The training data (790 notes) were distributed in two batches (an initial set of 521 notes, followed by a development set of 269 notes). While the initial training dataset was used for rule engineering and building of lexical resources, the development set was used for internal validation during the implementation.

Data	Micro		
	Precision	Recall	*F*-score
Initial training set (521 notes)	85.64	92.63	89.01
Development set (269 notes)	83.88	91.84	87.68
Evaluation set (514 notes)	85.57	90.07	87.76

attribute appears to be justified. Although the number of errors generated was not large, more sophisticated temporal information [23] could have contributed to the increase of the system's performance.

While the design and implementation of rule-based systems is known to be time consuming, in this case the whole system was engineered within ~6 weeks FTE (full-time equivalent), with the system fully operational within a month with further tests aiming to improve its efficiency in the remaining time. We have combined different expertise within the team, covering both clinical aspects and text mining experience, which allowed for a rapid domain-driven development of lexicalized rules. We purposely separated designing the common syntactical patterns for the identification of risk factors from the lexical modelling; therefore, the system can be tailored for the recognition of other targeted mentions by providing the necessary vocabularies, possibly from the existing clinical terminologies. Nonetheless, a significant number of rules involved the identification and "interpretation" of specific (measured) indicators (e.g. "LDL 111" is an indicator of high cholesterol); such rules will require redevelopment in case of a new task and potential linking to a clinical knowledge base.

Table 6
Results per risk factor class in the evaluation set.

Class	Frequency	Micro		
		Precision	Recall	*F*-score
Obesity	262	83.15	86.64	84.66
Diabetes	1189	93.27	79.83	86.03
CAD	1021	78.11	69.64	73.63
Hypertension	1308	95.53	85.92	90.47
Hyperlipidemia	751	90.94	82.82	86.69
Family history	514	95.91	95.91	95.91
Smoker	514	85.21	85.55	85.38
Medication	5825	82.24	96.97	89.00
Overall (run 1)	11,384	85.57	90.07	87.76

5. Conclusions

The objective of the i2b2 2014 task was to recognize heart disease risk factors from clinical narratives of diabetic patients and

assign the respective time intervals. In this paper we have described a methodology that is based on local rules lexicalized with extensive vocabularies that represent specific classes. The mention-level results were aggregated at the document level. The time attribute for each class relied on a number of specific default values. The overall performance of 88% *F*-score suggests that a lexicalized rule-based approach combined with default values can be used to process clinical notes to identify risk factors and monitor progression of heart disease on a large-scale, providing necessary data for clinical and epidemiological studies.

Future work includes the implementation of temporal extraction that will assist in the assignment of time values. Identification of a wider context of a disease or medication mention (e.g. relevant section (history, directions, course of treatment, allergies) and whether the mention refers to an event that is questionable/planned/negated are other areas that can contribute to better system performance. Finally, adding a clinical knowledge base and, in a real-world settings, the use of structured data (e.g. test/laboratory results) that appear in EHRs is a potential approach that can be used for data integration, validation and consolidation.

Availability

The resources developed as part of the system are available at http://gnode1.mib.man.ac.uk/tools/i2b2-2014-task2/.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This work has been supported by Health e-Research Centre (HeRC), The Christie Hospital NHS Foundation Trust, the Kidscan charity, the Royal Manchester Children's Hospital and the Serbian Ministry of Education and Science (projects III44006; IIII47003).

References

[1] World Health Organization. The Top 10 Causes of Death. <http://www.who.int/mediacentre/factsheets/fs310/en/>.
[2] A.S.V. Shah, M. Griffiths, K.L. Ken, D.A. McAllister, A.L. Hunter, A.V. Ferry, A. Cruikshank, et al., High sensitivity cardiac troponin and the under-diagnosis of myocardial infarction in women: prospective cohort study, BMJ 350 (2015) g7873.
[3] A. Stubbs, C. Kotfila, H. Xu, Ö. Uzuner, Practical Applications for NLP in Clinical Research: the 2014 i2b2/UTHealth Shared Tasks, J. Biomed. Inform 58S (2015) S1–S5.
[4] C. Friedman, L. Shagina, Y.A. Lussier, G. Hripcsak, Automated encoding of clinical documents based on natural language processing, J. Am. Med. Inform. Assoc. 11 (5) (2004) 392–402 [Epub 2004 June 7].
[5] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kiper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.
[6] I. Spasić, J. Livsey, J.A. Keane, G. Nenadic, Text mining of cancer-related information: review of current status and future directions, Int. J. Med. Inform. 83 (9) (2014) 605–623, http://dx.doi.org/10.1016/j.ijmedinf.2014.06.009.
[7] S. Sohn, J.P.A. Kocher, C.G. Chute, G.K. Savova, Drug side effect extraction from clinical narratives of psychiatry and psychology patients, J. Am. Med. Inform. Assoc. 18 (Suppl 1) (2011) i144–i149.
[8] S. Goryachev, K. Hyeoneui, Z.T. Qing, Identification and extraction of family history information from clinical reports, in: AMIA Annual Symposium Proceedings, vol. 2008, American Medical Informatics Association, 2008.
[9] Y. Wang, Annotating and recognising named entities in clinical notes, in: Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, Association for Computational Linguistics, 2009.
[10] J. Patrick, Li. Min, High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge, J. Am. Med. Inform. Assoc. 17 (2010) (2009) 524–527.
[11] I. Spasic, F. Sarafraz, A.J. Keane, G. Nenadic, Medication information extraction with linguistic pattern matching and semantic rules, J. Am. Med. Inform. Assoc. 17 (2010) 532–535.
[12] H. Yang, Automatic extraction of medication information from medical discharge summaries, J. Am. Med. Inform. Assoc. 17 (2010) 545–548.
[13] B. Rink, S. Harabagiu, R. Kirk, Automatic extraction of relations between medical concepts in clinical texts, J. Am. Med. Inform. Assoc. 18 (5) (2011) 594–600.
[14] S. Jonnalagadda, T. Cohen, S. Wu, G. Gonzalez, Enhancing clinical concept extraction with distributional semantics, J. Biomed. Inform. 45 (1) (2012) 129–140.
[15] Y. Xu, K. Hong, J. Tsujii, E. I-Chao Chang, Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries, J. Am. Med. Inform. Assoc. 19 (2012) 824–832, http://dx.doi.org/10.1136/amiajnl-2011-000776.
[16] M. Fiszman, G. Rosemblat, C.B. Ahlers, T.C. Rindflesch, Identifying risk factors for metabolic syndrome in biomedical text, in: AMIA Annual Symposium Proceedings, Vol. 2007, American Medical Informatics Association, 2007.
[17] Ö. Uzuner, I. Solti, E. Cadag, Extracting medication information from clinical text, J. Am. Med. Inform. Assoc. 17 (5) (2010) 514–518.
[18] S. Doan, N. Colier, H. Xu, H.P. Duy, M.T. Phuong, Recognition of medication information from discharge summaries using ensembles of classifiers, BMC Med. Inform. Decis. Mak. 12 (2012) 36.
[19] L. Deleger, C. Grouin, P. Zweigenbaum, Extracting medication information from narrative patient records: the case of medication-related information, J. Am. Med. Inform. Assoc. 17 (5) (2010) 555–558.
[20] Ö. Uzuner, Recognizing obesity and comorbidities in sparse data, J. Am. Med. Inform. Assoc. 16 (4) (2009) 561–570, http://dx.doi.org/10.1197/jamia.M3115.
[21] UMLS, 2014. <http://www.nlm.nih.gov/research/umls/>.
[22] W.W. Cohen, MinorThird: Methods for Identifying Names and Ontological Relations in Text Using Heuristics for Inducing Regularities from Data, 2004. <http://github.com/TeamCohen/MinorThird/>.
[23] A. Kovacevic, A. Dehghan, M. Filannino, J. Keane, G. Nenadic, Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives, J. Am. Med. Inform. Assoc. http://dx.doi.org/10.1136/amiajnl-2013-00.