Journal of Biomedical Informatics 58 (2015) S133–S142

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal relations from clinical records

Cyril Grouin [a,*], Véronique Moriceau [a,b], Pierre Zweigenbaum [a]

[a] LIMSI–CNRS, Orsay, France
[b] Université Paris-Sud, Orsay, France

ARTICLE INFO

Article history:
Received 14 May 2015
Revised 16 June 2015
Accepted 22 June 2015
Available online 2 July 2015

Keywords:
Natural language processing
Electronic health records
Risk factors
Program evaluation

ABSTRACT

Background: The determination of risk factors and their temporal relations in natural language patient records is a complex task which has been addressed in the i2b2/UTHealth 2014 shared task. In this context, in most systems it was broadly decomposed into two sub-tasks implemented by two components: entity detection, and temporal relation determination. Task-level ("black box") evaluation is relevant for the final clinical application, whereas component-level evaluation ("glass box") is important for system development and progress monitoring. Unfortunately, because of the interaction between entity representation and temporal relation representation, glass box and black box evaluation cannot be managed straightforwardly at the same time in the setting of the i2b2/UTHealth 2014 task, making it difficult to assess reliably the relative performance and contribution of the individual components to the overall task. *Objective:* To identify obstacles and propose methods to cope with this difficulty, and illustrate them through experiments on the i2b2/UTHealth 2014 dataset. *Methods:* We outline several solutions to this problem and examine their requirements in terms of adequacy for component-level and task-level evaluation and of changes to the task framework. We select the solution which requires the least modifications to the i2b2 evaluation framework and illustrate it with our system. This system identifies risk factor mentions with a CRF system complemented by hand-designed patterns, identifies and normalizes temporal expressions through a tailored version of the Heideltime tool, and determines temporal relations of each risk factor with a One Rule classifier. *Results:* Giving a fixed value to the temporal attribute in risk factor identification proved to be the simplest way to evaluate the risk factor detection component independently. This evaluation method enabled us to identify the risk factor detection component as most contributing to the false negatives and false positives of the global system. This led us to redirect further effort to this component, focusing on medication detection, with gains of 7 to 20 recall points and of 3 to 6 F-measure points depending on the corpus and evaluation. *Conclusion:* We proposed a method to achieve a clearer glass box evaluation of risk factor detection and temporal relation detection in clinical texts, which can provide an example to help system development in similar tasks. This glass box evaluation was instrumental in refocusing our efforts and obtaining substantial improvements in risk factor detection.

1. Introduction

Medical records for diabetic patients contain information about heart disease risk factors. In electronic health records, this information is mainly given in the form of unstructured text. To improve patient care, automatic extraction of medically relevant information can provide clinicians with clues on diverse heart disease risk factors, and their progression over time. Tracking the progression over time of heart disease risk factors in diabetic patients was the topic of the i2b2/UTHealth 2014 challenge [1,2]. The determination of risk factors from clinical texts requires to detect diseases (*diabetes, coronary artery disease*), associated risk factors (*cholesterol and hyperlipidemia, hypertension, obesity, smoker status, family history*), and clues thereof (*medications*); the other part of the task demands to find where in time most of these risk factors occurred on the patient's timeline.

* Corresponding author.
E-mail address: cyril.grouin@limsi.fr (C. Grouin).

http://dx.doi.org/10.1016/j.jbi.2015.06.014
1532-0464/

This task description led us to split our system into two components: one for risk factor detection (possibly decomposed into as many sub-components as types of risk factors), and one for temporal relation determination. Combined in a pipeline, they enumerate the risk factors present in a patient record then compute their temporal relations to the current visit. Overall system performance is indeed the most important type of evaluation for the final clinical task. However, when this evaluation reveals a certain number of false positives or false negatives, it is also important to know which component most needs improving. Principled system development should therefore provide a way to evaluate each component independently of each other and of the full system, ideally in such a way as to predict their impact on overall system performance.

We show in this paper that this is not straightforward to obtain in the i2b2/UTHealth 2014 challenge risk factors task (Section 3), and explain why. We examine potential solutions to this problem, find out that none is fully satisfactory, and implement the one which requires the least modification to the i2b2 evaluation framework (Section 4). We illustrate its application on our risk factor and temporal relation detection system (Sections 5 and 6) and use it to point more clearly at directions for its improvement. We follow the most promising of these directions and obtain substantial gains in system performance (Section 7), then conclude (Section 8).

2. Related work

Information extraction tasks often proposed dual evaluation scenarios in which both full-task (black box) evaluation and component (glass box) evaluation were organized. This is often non-trivial to achieve because of interrelationships between components. For example, the detection of relations generally depends on the former detection of entities which these relations link (note that joint methods are also proposed by some authors, but are not the subject of this paper).

Binary semantic relations such as those which hold between medical problems, tests, and treatments [3] rely on the detection of these concept types. Nevertheless, the 2010 i2b2/VA challenge defined and evaluated two separate sub-tasks through micro-averaged precision, recall and F-measure: concept extraction and relation classification. This provided a glass box evaluation of each sub-task. It did not propose an evaluation of end-to-end concept extraction and relation classification systems, but this (black-box) evaluation would have been easy to run based on the evaluation measures of the relation extraction sub-task.

Binary temporal relations (before, after, etc.) which link events and temporal expressions depend on the detection of these events and times. The 2012 i2b2 temporal relations challenge [4] defined sub-tasks for the identification of EVENTs, the identification of temporal expressions (TIMEX3s), and the detection of the temporal relations between them. This led participants to create three separate components and enabled them to evaluate each of those components through glass box evaluation. Non-trivial issues stemmed from the need to normalize various equivalent configurations of temporal relations. For this purpose their transitive closure was computed before computing their F-measure. Note that choosing the transitive closure instead of, e.g., a minimal underlying temporal graph [5], changes the number of relations that are evaluated.

Co-reference relations detect which mentions in a text refer to the same entities; therefore the determination of these relations also depends on the detection of entity mentions [6]. The 2011 i2b2/VA challenge [7] defined separate sub-tasks for mention detection and co-reference resolution, thus providing glass box evaluation for each sub-task. It also defined an end-to-end task where system mentions were used as input to the co-reference resolution step. Co-reference resolution was evaluated through the MUC, B^3, and CEAF metrics. However, Cai and Strube [6] showed that the original B^3 and CEAF measures have problems when applied to end-to-end systems, i.e., with concept mentions computed by a first component, and hence not always adequate; they proposed adaptations of these measures to alleviate these problems.

Sometimes relations are viewed instead as concept attributes: this was the case of the 2014 i2b2/UTHealth challenge [2], which defined a task where risk factors had to be detected, together with their temporal relation to the document creation time (DCT). This challenge relates to a large subset of the history of previous i2b2 challenge tasks as well as to the 2014 ShARe/CLEF eHealth T2 shared task [8]. This challenge, which we describe in more detail in Section 3, only defined a black-box evaluation of the end-to-end task, but did not provide a separate, glass-box evaluation of risk factor detection and temporal relation detection. To develop an optimal end-to-end system, we considered it important to obtain a separate evaluation for each of these components. We present in Section 4 the issues we encountered when trying to obtain such a glass-box evaluation and the solution we adopted. We illustrate this glass-box evaluation with the system we developed for the i2b2/UTHealth 2014 challenge (Sections 5 and 6) and discuss how it helped focus error analysis and system improvement (Section 7).

3. Definition of the i2b2/UTHealth 2014 composite task

3.1. Corpus

The corpus we used for the following experiments is the 2014 i2b2/UTHealth corpus, composed of 1304 patient records from 3 cohorts of diabetic patients for a total of 296 patients. For each patient, about 3 to 5 records are provided per patient, referring to different times in the patient's timeline. The training corpus contained 790 records (178 patients) and the test corpus contained 514 records (118 patients). Patients were distinct between training and test corpora. Based on a random selection, we split the training corpus into our training sub-corpus (89 patients, 390 records) to develop our system, and our development sub-corpus (30 patients, 131 records) to tune the system. Our internal test sub-corpus is composed of 269 records (59 patients).

In our experiments, results on the internal test sub-corpus were obtained with systems trained on the training + development sub-corpora, and results on the official test corpus were obtained with systems trained on the full training corpus.

3.2. Task description: risk factor detection

The task consists in identifying risk factors for diabetic patients in clinical records [9] among 8 categories: diabetes mellitus (DM), coronary artery disease (CAD), hyperlipidemia (HLD), hypertension (HTN), medication (MED), obesity (OBE), family history of CAD (FAM), and smoker status (SMO). The first six categories are events which may take place before, during or after the current visit. Information on how risk factor events are expressed in the document must be specified: for instance, "HTN" or "hypertension" are explicit *mentions* of the risk factor, whereas a test result such as a blood pressure measurement over 140/90 mm/hg is categorized as a *high bp*. This defines sub-types of risk factor events, the full set of which is shown in Table 1. For instance, an expression such as "150/90" should be recorded as *HTN* with sub-type *high bp*.

The task is a document-level entity detection task: what must be determined is whether a risk factor of a given sub-type is present or not in a document, not its specific occurrences and