Journal of Biomedical Informatics 53 (2015) 36–48

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A natural language processing pipeline for pairing measurements uniquely across free-text CT reports

Merlijn Sevenster [a,*], Jeffrey Bozeman [b], Andrea Cowhy [b], William Trost [b]

[a] Clinical Informatics, Interventional & Translational Solutions, Philips Research North America, 345 Scarborough Road, Briarcliff Manor, NY, USA
[b] Department of Medicine, University of Chicago, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 14 January 2014
Accepted 30 August 2014
Available online 6 September 2014

Keywords:
Natural language processing
RECIST
Radiology report
Oncologic measurement
Information correlation

ABSTRACT

Objective: To standardize and objectivize treatment response assessment in oncology, guidelines have been proposed that are driven by radiological measurements, which are typically communicated in free-text reports defying automated processing. We study through inter-annotator agreement and natural language processing (NLP) algorithm development the task of pairing measurements that quantify the same finding across consecutive radiology reports, such that each measurement is paired with at most one other ("partial uniqueness").

Methods and materials: Ground truth is created based on 283 abdomen and 311 chest CT reports of 50 patients each. A pre-processing engine segments reports and extracts measurements. Thirteen features are developed based on volumetric similarity between measurements, semantic similarity between their respective narrative contexts and structural properties of their report positions. A Random Forest classifier (RF) integrates all features. A "mutual best match" (MBM) post-processor ensures partial uniqueness.

Results: In an end-to-end evaluation, RF has precision 0.841, recall 0.807, *F*-measure 0.824 and AUC 0.971; with MBM, which performs above chance level ($P < 0.001$), it has precision 0.899, recall 0.776, *F*-measure 0.833 and AUC 0.935. RF (RF + MBM) has error-free performance on 52.7% (57.4%) of report pairs.

Discussion: Inter-annotator agreement of three domain specialists with the ground truth ($\kappa > 0.960$) indicates that the task is well defined. Domain properties and inter-section differences are discussed to explain superior performance in abdomen. Enforcing partial uniqueness has mixed but minor effects on performance.

Conclusion: A combined machine learning–filtering approach is proposed for pairing measurements, which can support prospective (supporting treatment response assessment) and retrospective purposes (data mining).

1. Introduction

1.1. Background and motivation

Health care enterprises are under continuous pressure to produce "predictable and reproducible outcomes" from clinical examinations and diagnostic tests, which "requires that diagnostic information be expressed in quantitative form" [1]. In oncology, guidelines have been proposed to standardize and objectivize treatment response assessment, such as the World Health Organization guidelines [2] and RECIST (Response Evaluation Criteria in Solid Tumors) [3]. These guidelines are primarily based on radiologic measurements of selected index lesions [4].

Lesion measurements are generally made by radiologists [5] after selecting a set of representative and/or previously measured lesions. They are subsequently communicated by means of free-text radiology reports [6,7]. The free-text nature of radiology reports prohibits automated processing of their information content in support of downstream consumers [7–10], such as oncologists and clinical research associates (CRAs). Oncologists rely on reported measurements and qualitative assessments to synthesize treatment response status and to determine an optimal care plan. CRAs parse radiology reports of cancer patients to transcribe their lesion measurements into clinical trial databases.

If lesion measurement data were available in structured [11] and digital form [12], as a supplement to the narrative radiology report, it could be leveraged to support downstream consumers. Properly grouped by lesion, structured measurement data could be used to effortlessly compute RECIST scores and could be

* Corresponding author. Fax: +1 (914) 945 6330.
E-mail address: Merlijn.sevenster@philips.com (M. Sevenster).

http://dx.doi.org/10.1016/j.jbi.2014.08.015
1532-0464/

inserted automatically into clinical trial databases. Oncology information systems that accomplish this have the potential to minimize transcription errors [13,14], improve efficiency and facilitate data-driven treatment response assessment for on- and off-trial cancer patients alike. They may further open up novel application areas such knowledge discovery through data mining [15], cohort selection using advanced queries [16], and multi-disciplinary collaboration in oncology [17].

Such oncology information systems face three technological challenges. (1) Data acquisition: Data elements are obtained from structured or narrative sources [18]. In the latter case, pertinent data elements can be disclosed by natural language processing (NLP) techniques [19], for instance, for automatically synthesizing treatment histories [20] or populating registries of cancer patients [21]. (2) Data integration: multi-source and longitudinal data elements are mapped into one coherent data structure [22,23]. (3) Data presentation: integrated data elements are presented graphically to the user [24,25].

Systems that address these challenges in isolation have been reported more frequently in the literature than systems that address them in combination. A recent system that exemplifies the latter category extracts neuro-oncologic findings from a history of radiology reports and normalizes it with respect to a controlled interval change vocabulary containing, e.g., "existing" and "improving" [26,27].

1.2. Task definition

In this work, we introduce the task of extracting and pairing measurements across consecutive reports. A unique feature of the task is that across two consecutive reports, a measurement is paired with at most one other measurement. We call this the *partial uniqueness* condition. This condition is motivated by the observation that in clinical practice once measured the vast majority of lesions continues to be measured in subsequent follow-up exams, unless the lesion resolves or if the radiologist fails to report its measurement. The output of automated solvers of this task can be utilized by downstream modules, e.g., for visualization or automated treatment response assessment.

In this paper, we propose a natural language processing (NLP) pipeline that consumes a patient's history of narrative radiology reports and segments [28] them in the pre-processing phase. Then, addressing challenge 1, measurements are extracted, normalized and labeled with respect to their temporal orientation [29]. Finally, addressing challenge 2, measurements are paired across reports and a filter is proposed that enforces the partial uniqueness condition, which, as we argue above, holds for the vast majority of lesions.

1.3. Related work

All components in the pipeline proposed in this work are home grown, leveraging the results of prior research projects. Third-party engines can, however, be used to achieve parts of the aimed measurement pairing functionality.

Report segmentation, i.e., the automated break down of a medical narrative document in its main components (e.g., sections, subsections and sentences) has been studied in the literature, either as component of a general-purpose system (e.g., MedLEE [30], Leximer [31] and cTAKES [32]) or as a dedicated engines [28]. A potential downside of general-purpose systems is that their respective output must be processed further to retrieve the additional radiology-specific structure that cannot be assumed to exist in narrative documents from other medical specialties (e.g., oncology notes) that are within the scope of the general-purpose system. MedLEE recognizes measurements, which constitute the core tokens in the measurement matching task. This engine can thus be used as an alternative to our measurement extraction engine.

In previous research, we developed a pipeline that extracts and normalizes measurements from radiology reports. In addition, a classification engine in this pipeline was developed that detects the "temporal orientation" of a given measurement, that is, if the measurement was made on the current or prior exam. This engine was deployed to estimate the number of measurements across radiology reports of different modalities and anatomies [33]. To the best of our knowledge such methods have not been researched before. Indeed, we are not aware of any information extraction system that produces an output from which a measurement's temporal orientation can be derived with relatively lightweight logic.

The work presented in this paper is an extension of a conference paper [34] in the sense that it includes chest reports in its ground truth in addition to the initial abdomen reports. Further, we extended the pipeline with the aforementioned MBM engine and report on micro analysis results.

2. Methods and materials

We explore two approaches to automatically pairing measurements, which we define as a binary classification problem of instances. In the context of two consecutive reports, an *instance* is a pair of measurements from the Findings sections of the prior and current report, respectively. An instance is *positive* or a *match*, if its measurements quantify the same clinical finding on their respective exams [34], see Fig. 1, which will serve as a running example throughout this section. Measurements from non-Findings sections are excluded as they report slice thickness (Technique) or re-iterate measurements from the Findings sections as a means to support the overall impressions of the radiological examination (Conclusion).

The first approach uses machine-learning methods to integrate features that quantify volumetric similarity between measurements, semantic similarity between their respective narrative contexts and structural properties of the measurements' report positions. The second extends the first approach by a novel post-processing technique based on *mutual best matches*, which enforces the partial uniqueness condition. The proposed pipeline, including the post-processing filter, is schematically displayed in Fig. 2.

A ground truth is constructed based on the abdomen and chest CT reports of 50 patients each. The ground truth's quality and reproducibility of the ground truth construction process are assessed in an inter-annotator agreement study with three CRA domain specialists. The performance of the entire pipeline is assessed in an end-to-end evaluation with and without the mutual best match filter against the ground truth.

2.1. Ground truth development

2.1.1. Corpus

A database of radiology reports was obtained from The University of Chicago Medical Center. The reports were authored using dictation software (PowerScribe, Nuance, current version 3.0.19.6) with in-house developed reporting templates with all-caps section headers and anatomical paragraph headers, see Fig. 1.

We de-identified our dataset using the following approach. All dates in the database, in the form of metadata as well as narrative references in the reports, were offset by randomly generated, patient-specific integers. All other types of HIPAA patient health information were removed using a homegrown engine driven by a collection of regular expressions. The database was accessed under waived IRB 13-0379.