Journal of Biomedical Informatics 62 (2016) 78–89

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports

Wei Wang [a], Kory Kreimeyer [b], Emily Jane Woo [b], Robert Ball [c], Matthew Foster [b], Abhishek Pandey [b], John Scott [b], Taxiarchis Botsis [b,*]

[a] Engility Corporation, Chantilly, VA, United States
[b] Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States
[c] Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States

ARTICLE INFO

Article history:
Received 14 January 2016
Revised 11 June 2016
Accepted 17 June 2016
Available online 17 June 2016

Keywords:
Natural language processing
Post-marketing surveillance
Temporal information

ABSTRACT

The sheer volume of textual information that needs to be reviewed and analyzed in many clinical settings requires the automated retrieval of key clinical and temporal information. The existing natural language processing systems are often challenged by the low quality of clinical texts and do not demonstrate the required performance. In this study, we focus on medical product safety report narratives and investigate the association of the clinical events with appropriate time information. We developed a novel algorithm for tagging and extracting temporal information from the narratives, and associating it with related events. The proposed algorithm minimizes the performance dependency on text quality by relying only on shallow syntactic information and primitive properties of the extracted event and time entities. We demonstrated the effectiveness of the proposed algorithm by evaluating its tagging and time assignment capabilities on 140 randomly selected reports from the US Vaccine Adverse Event Reporting System (VAERS) and the FDA (Food and Drug Administration) Adverse Event Reporting System (FAERS). We compared the performance of our tagger with the SUTime and HeidelTime taggers, and our algorithm's event-time associations with the Temporal Awareness and Reasoning Systems for Question Interpretation (TARSQI). We further evaluated the ability of our algorithm to correctly identify the time information for the events in the 2012 Informatics for Integrating Biology and the Bedside (i2b2) Challenge corpus. For the time tagging task, our algorithm performed better than the SUTime and the HeidelTime taggers (F-measure in VAERS and FAERS: Our algorithm: 0.86 and 0.88, SUTime: 0.77 and 0.74, and HeidelTime 0.75 and 0.42, respectively). In the event-time association task, our algorithm assigned an inappropriate timestamp for 25% of the events, while the TARSQI toolkit demonstrated a considerably lower performance, assigning inappropriate timestamps in 61.5% of the same events. Our algorithm also supported the correct calculation of 69% of the event relations to the section time in the i2b2 testing set.

Published by Elsevier Inc.

1. Introduction

The development of Natural Language Processing (NLP) systems for clinical texts has focused primarily on the extraction of clinical events and, secondarily, on their association with temporal information. The latter was specifically tackled in the 2012 Informatics for Integrating Biology and the Bedside (i2b2) Shared-Task and Workshop on Challenges in Natural Language Processing for Clinical Data, where participants were invited to develop systems for processing temporal relations in clinical records. According to Uzuner et al., "the 2012 i2b2 Challenge systems only scratched the surface in this task" and "open questions remain about the applicability of the developed systems for real life practical questions" [1].

The US Vaccine Adverse Event Reporting System (VAERS) is a Spontaneous Reporting System (SRS) that collects safety reports for adverse events (AEs) following immunization for vaccines licensed for use in the United States [2]. The US FDA (Food and Drug Administration) Adverse Event Reporting System (FAERS) is a database that contains information on adverse event and medication error reports submitted to FDA [3]. Reports submitted

* Corresponding author at: Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, FDA, 10903 New Hampshire Ave, WO71 – 1233, Silver Spring, MD 20993-0002, United States.
E-mail address: Taxiarchis.Botsis@fda.hhs.gov (T. Botsis).

http://dx.doi.org/10.1016/j.jbi.2016.06.006
1532-0464/Published by Elsevier Inc.

to the FDA from manufacturers, health care providers, and the public include AE narratives of varying length and text quality. They also provide a good example of real-life data that would benefit from NLP applications. Spontaneous post-marketing reports may provide early evidence of new and unexpected AEs that were not identified during premarketing clinical trials. As of April 2015, the VAERS contains more than 500,000 reports of AEs following immunization, and the FAERS contains more than 10 million reports of AEs after drug exposure.[1] Duggirala et al. estimated that VAERS and FAERS receive 35–40,000 and 770,000 reports per year, respectively, and this number is expected to grow [4]. The volume of textual information in these surveillance databases points to the need for NLP approaches to retrieve the key clinical and temporal information to aid medical and epidemiological review.

Many systems have been developed to respond to particular challenges requiring the processing of temporal information in clinical texts [5]. Rule-based, machine learning (ML), and hybrid systems have been developed to detect temporal relations [6–11]. Rule-based approaches primarily use domain-specific handcrafted rules that require a great amount of human effort to develop and lack portability to other domains [12]. ML approaches learn from annotated data sets and therefore performance often decreases when transferring a ML-based system from one domain to another [12]. In both approaches, temporal information extraction algorithms typically rely on grammatical and syntactical attributes, such as part-of-speech (POS) tags, event modalities, and tenses. For example, the baseline system used by D'Souza et al. relied on 40 grammatical features, 10 entity attributes and 7 semantic features [7]. The extraction of these attributes is problematic in low quality texts (poor grammar, many abbreviations, etc.), as found in many clinical narratives (VAERS and FAERS reports often fall into this category), and will considerably affect the performance of systems that rely on these attributes. It is therefore necessary to develop strategies with less dependency on the text quality.

Previous studies attempted to address the shortcomings in clinical time extraction using the Time Markup Language (TimeML) specification as the starting point [13]. Notable exceptions include the work of Zhou et al. on a Simple Temporal Problem Approach [14], and the Clinical Narrative Temporal Relation Ontology, which has been developed as a Web Ontology Language [15]. Additionally, Raghavan et al. have shown that a good deal of temporal information can be captured into useful timelines simply by assigning events into coarse time-bins [16]. However, it is the TimeML-based approaches that have remained popular in recent years. The THYME project is a large effort to both define a TimeML-based markup for clinical text and create a useful corpus of temporally marked-up electronic health record data using the TimeML specification [17]. A portion of the THYME corpus was recently used for the Clinical TempEval task at SemEval 2015. The participating teams were able to extract event and time features with high reliability, but the mapping of temporal relations suffered from poorer performance [18]. The 2012 i2b2 Challenge also focused on the temporal relations in clinical narratives and the eighteen teams that participated in the competition achieved encouraging performance [19,20]. After both challenges, Lin et al. evaluated their own multilayered temporal system against TempEval's THYME dataset and the i2b2 corpus and found superior performance; however, they still struggled with temporal mappings for long-distance relationships, such as cross-sentence relations [21].

[1] US VAERS and FAERS contained 526,991 and 10,146,489 reports, respectively, on 5/01/2015 (11:00 am).

We previously developed and evaluated a text mining system for the extraction of clinical features from VAERS reports [22,23]. The next generation of this system, the Event-based Text-mining of Health Electronic Records (ETHER), has been equipped with new functionalities and expanded to process FAERS reports. ETHER includes a novel rule-based algorithm that extracts temporal information and assigns it to the ETHER clinical features (or events). It consists of a time expression tagger and a module to assign the appropriate time to events. Our time tagger identifies a broad range of time expression patterns. The key contribution of our algorithm is its minimal dependency on text quality when extracting temporal associations from the narrative. In our approach both the event and time entities are characterized by their primitive properties only and no additional syntactic information, such as POS and tense, is required. Instead, properties like the text location of each event or time entity are used to support the identification of temporal associations. We particularly construct time impact zones for evaluating the time flow and organizing the temporal information within a narrative. The core algorithm attempts to assign a timestamp to all events with certain exceptions defined through secondary rules. These rules are domain-specific and support certain needs, e.g. the secondary rules in this work have been tailored for safety surveillance reports. These domain-specific rules are all incorporated in the preprocessing stage and act independently of the core algorithm. Therefore, the core algorithm can be potentially incorporated into any other NLP system and assign timestamps to completely different types of events in a different application. We here describe our algorithm, compare it with existing systems using a set of VAERS and FAERS reports, and evaluate its applicability to other settings using the 2012 i2b2 Challenge corpus [19,20].

2. Methodology

Temporal information extraction consists of the following subtasks: (1) event extraction; (2) time expression tagging; and (3) association of events and temporal information.

2.1. Event extraction

The event extraction process can vary significantly between different systems, depending on how an "event" is defined. For example, the organizers of the 2012 i2b2 Challenge defined four types of clinically relevant events: clinical concepts, clinical departments, evidentials, and occurrences [19]. In contrast, many systems define events as verbs and event extraction is based on verb features such as class, tense, aspect, and polarity. For example, Jung et al. focused on verbs and verb features for event extraction and then linked them to clinical concepts to construct a timeline [24]. In this study, we define events as ETHER clinical features, which are described below.

Our software, ETHER, extracts certain clinical features: (i) *primary diagnostic features*: primary diagnosis and cause of death; (ii) *secondary diagnostic features*: secondary diagnosis, rule-out diagnosis, and symptoms; (iii) *causal assessment features*: time to the onset of the first symptom(s) following exposure to a product, family and medical history; and (iv) *exposure-related features*: vaccine and drug names [23]. These clinical features will play the role of events and will be associated with time information that is captured and calculated by our time algorithm. A detailed description of the extraction of clinical features can be found in [23]. This paper focuses on the time expression tagging and the association of events and temporal information only.