A CNL-based Method for Detecting Disease Negation

Joan Byamugisha and **Nomonde Khalo**
IBM Research Africa
45 Juta Street, Braamfontein
Johannesburg, 2000, South Africa
joan.byamugisha@ibm.com and nomonde.khalo@ibm.com

Abstract

Negation detection is a key feature to the processing of biomedical text, and it involves two steps: identifying a medical term of interest in text and identifying that that medical term is mentioned as absent. However, processing biomedical text is made complex by the presence of medical jargon that typically requires custom systems, and detecting negation is complicated further because the representation of negation in natural language varies according to the grammar used. We investigated the use of a CNL with a general-purpose semantic parser to detect negation. Our CNL was created by representing medical terms as their semantic types and restricting the definition of the expression of negation. Through this method, we identified three kinds of negation–explicit negation, implicit negation, and explicit implicit negation. A pilot evaluation of our method on a sample of radiology reports achieved an F1 score of 0.99 on the sentences that could be parsed.

1 Introduction

Natural language (also referred to as free text or narrative) is the most wide-spread, comprehensive, and convenient medium for healthcare personnel to present medical information, for example, in patient progress notes, radiology and pathology reports, and discharge summaries (Wang et al., 2018). Narrative patient reports contain several medical concepts (naming entities such as body parts, drugs, symptoms, diseases, medical tests, and treatments) and the relations between the concepts (Wang et al., 2018). Identifying and distinguishing among different entities is the basis of biomedical NLP tools for text classification, named entity recognition, and text summarization, section detection, and negation detection, among others. In this paper, we focus on the task of disease negation detection, an important aspect in biomedical NLP, which involves two steps: identifying a medical term of interest in text and identifying that that medical term is mentioned as absent (Chapman et al., 2001). Both these tasks are complicated by the nature of medical jargon and the variation in the representation of negation in natural language.

Current approaches taken to solve the problem of identifying negated medical terms can be categorized as syntactic-based, ontology-based, and corpus-based. Syntactic-based systems typically rely on the use of custom regular expressions for pattern matching and to combine grammar parsing with standard expression matching (Huang and Lowe, 2007). Ontology-based systems, such as (Elkin et al., 2005), apply the knowledge from an ontology to standardize the representation of medical terms, thus enabling their automatic interpretation during negation detection. Corpus-based systems use machine learning algorithms to learn the scope of negation and treat negation detection as a classification task (Slater et al., 2021). The limitations of these approaches include: for syntactic-based systems, limited coverage due to being restricted to the syntax defined in a regular expression, offering little or no contextualization of the syntax to the semantics in the text, and tedious maintenance; for ontology-based systems, the level of semantic contextualization provided is limited and lacks broader coverage of the entire text; and for corpus-based systems, they require large amounts of data that are not readily available in the healthcare domain; and the black-box nature of machine learning algorithms makes it impossible to assess what aspects of negation they are learning.

On the other hand, there exist very efficient and highly accurate general-purpose semantic parsers that can be used to detect negation. The problem here is that the presence of medical jargon in natural language text increases the complexity and ambiguity already inherent in natural language, and renders attempts at using general-purpose parsers unreliable, as they result in inaccurate semantic representations. For example, parsing sentences with medical jargon using the ACE parser (Packard, 2013) produces results with semantic categories identified as ‘unknown’. As the source of the complexity is the presence of medical jargon, we hypothesized that reducing this complexity to a vocabulary that can be parsed by general-purpose semantic parsers can reduce the negation detection problem to that present in a domain-independent vocabulary, and enable the use of a general-purpose semantic parser to perform negation detection of disease entities. Our work contributes: (1) a method of defining a CNL by first looking at natural language and then restricting its lexicon and expression of negation; and (2) a pilot method, still to be evaluated comprehensively, for negation detection in a medical domain using a general-purpose English parser.

2 CNL-based Negation Detection

For the task of negation detection, we sought to limit the impact of medical jargon by converting medical text into a restricted version that can be parsed deterministically, CNL, using a general-purpose English parser. We used an efficient linguistic processor for Head-driven Phrase Structure Grammars (HPSGs), the Answer Constrained Engine (ACE) (Packard, 2013), which supports most modern computational linguistic features. A broad-coverage symbolic grammar of English–the English Resource Grammar (ERG) (Copestake and Flickinger, 2000), was used to parse the CNL and analyzed the Minimal recursion semantics (MRS) (Copestake et al., 2005) representations for signals of the negation of mentioned entities. The following sections present details on the materials, methods, and results of this work.

2.1 Materials

The data used in this investigation was obtained from the corpus of radiology reports from the Mimic CXR dataset (Johnson et al., 2019). We used a sample of 100 reports and considered only the sentences associated with the sections in a report which contain conclusions about the findings in a report. These sections are labeled as *FINDINGS*, *IMPRESSIONS*, or *CONCLUSIONS*, or their singular forms. Some reports do not have these sections, while others have at least one of these sections. From our sample of 100 reports, 92 were found to possess at least one of these sections, and from these reports, 345 sentences were obtained. These sentences were examined manually to remove any sentence that contained the after effects of report deidentification (such as ‘___ at ___ on ___.’ and ‘Analysis is performed in direct comparison with the next preceding similar study of ___.’). These were removed, resulting in a final dataset of 316 sentences.

A ground-truth dataset was created manually for these sentences. The criterion used when labeling the dataset was that, if there is at least one indicator of disease which is mentioned as present, then that sentence is labeled as *N* for ‘not negated’, otherwise, it is labeled as *Y*. The rationale behind this labeling scheme is that the purpose of negation detection is to identify patients who have at least one disease indicator as opposed to patients who have none. Therefore, sentences such as, ‘Bilateral pleural effusions, severe pulmonary edema, cannot exclude pneumonia.’ and ‘The heart size is normal, but the pulmonary vasculature is still mildly engorged.’ are labeled as *N* (not negated); while a sentence such as ‘Heart size is enlarged but stable.’ is labeled as *Y* (negation present). Of the 316 sentences in the dataset, 136 sentences were annotated as negating the mentioned disease indicators, while 180 were annotated as possessing at least one present disease indicator.

Our CNL was created by representing medical terms as their semantic types and restricting the definition of the expression of negation. Though simple, we ensured that the result possessed all the four properties by which a language can be regarded as a CNL: (1) it is based on exactly one natural language, its base language; (2) the more restrictive lexicon is the most important difference between it and its base language, and we restrict further its expression of negation; (3) it preserves most of the natural properties of its base language; and (4) it is explicitly and consciously defined (Kuhn, 2014). When constructing our CNL, we selected two semantic types–*Anatomy* and *Disease*–because we are focusing on detecting the negation of diseases. The *Anatomy* semantic type is required in order to contextualize disease mentions that are expressed through an anatomical region where a disease occurs. Creating the CNL requires identifying a medical term in text and then determining its semantic type. We relied on the knowledge in the Unified Medical Language System (UMLS) to determine whether a medical term represents a disease or an anatomy, and we selected six terminologies in the 2020 release of the UMLS metathesaurus purposively so as to have a broad coverage with which to identify diseases and anatomies. For parsing with ACE (Packard, 2013), we used the English Resource Grammar (ERG) (Copestake and Flickinger, 2000) and selected MRS (Copestake et al., 2005) as the representations with which to analyze the results.

2.2 Methods

First, QuickUMLS (Soldaini and Goharian, 2016) was used on each sentence to extract medical concepts and their corresponding Concept Unique ID (CUI). 392 medical terms and their CUIs were extracted from 316 sentences. Next, the semantic type of an entity was determined by mapping a CUI to each of the five selected terminologies, which is possible because the UMLS Facilitates conceptual mappings among terminologies. If a mapping from a source terminology to a target terminology produces concept(s), then it implies that that concept is found in the target terminology, and is, therefore, of the semantic type represented by that terminology. Based on this criterion, of the 392 medical entities extracted, 62 entities representing anatomies and 64 entities representing diseases were found.

After this, creating a CNL of each sentence was done by replacing a medical term with the semantic type associated with it. For example, 'right rotator cuff' becomes *Anatomy* and 'interstitial edema' becomes *Disease*. Additionally, where multiples of the same semantic type are present in a sentence, they are numbered so as to differentiate them to the parser and maintain the semantics in a sentence. For example, 'There is no evidence of pneumothorax, pleural effusion, pulmonary edema, or pneumonia.' becomes, 'There is no evidence of Disease1, Disease2, Disease3, or Disease4.'. Through this process, medical jargon is reduced to representations of proper nouns that can be parsed using ACE. Finally, we applied two types of negation: explicit negation and implicit negation. Explicit negation is detected through the presence of negation markers and qualifiers in the MRS output. In MRS, explicit negation is represented with '*neg*'. Additionally, the quantifier for a noun, if found to be '*no*', semantically signifies that an entity is present zero times, hence, negation. For example, in the sentence 'There is no evidence of pneumothorax, pleural effusion, pulmonary edema, or pneumonia.', 'no' is a quantifier signifying zero 'evidence'; as opposed to, say, 'some evidence of' or 'only evidence of'. In our CNL, we restrict implicit negation as detected through a limited vocabulary. We extracted adjectives, nouns, and verbs from the MRS representations and identified semantic constructions that indicate the presence or absence of a disease. For the former, constructions such as 'present_a_1', 'indication_n_of', and 'worsen_v_cause' point to the presence of a disease; for the latter, constructions such as 'clear_a_of', 'normal_n_1', and 'rule_v_out' point to the absence of a disease.

2.3 Results

Of the 316 sentences with ground-truth negation values, 267 were parsed successfully with ACE, while 49 (15.51%) could not be parsed and had no MRS representations. We, therefore, present results obtained from the 267 sentences. 89 sentences (33.33%) were found to contain a conjunction, while 28 (10.49%) contained the explicit negation construct '*neg*' and 70 (26.22%) contained 'no' as a quantifier of diseases. Therefore, the total number of sentences with constructs associated with explicit negation was 98 (36.7%).

Of the 115 sentences annotated as negated in the ground-truth, 81 were identified correctly through explicit negation. The false negatives from using explicit negation only comprise sentences that either express a disease by referring to an anatomical region instead of a disease directly; or describe the absence of a disease without negating its presence explicitly, rather implicitly. Examples of sentences where a disease is negated by describing the affected anatomy are, 'The heart size remains normal as well as the thoracic aorta which follows the scoliotic curvature in its descending portion remains within normal limits.', and 'The cardiac, mediastinal and hilar contours appear stable.'. Cases where the presence of a disease is negated implicitly are, 'Since the prior exam, the lung volumes have improved.', and 'The right perihilar opacification and bilateral pleural effusions have resolved.'.

For implicit negation, we catered for two cases: (1) implicit negation either anatomically or through disease; and (2) negation of implicit negation. When investigating this, the following parts-of-speech were extracted: 155 adjectives, 166 nouns, and 91 verbs. Of these, 5 adjectives, 5 nouns, and 5 verbs were included in a vocabulary as signifying the absence of a disease; while 3 adjectives, 9 nouns, and 7 verbs were included in a vocabulary as signifying the presence of a disease. Implicit negation considers two kinds of semantics–those that signify the presence of a disease and those that signify the absence of a disease. The vocabulary required to identify implicit negation is very small when compared to the parts-of-speech extracted, that is, 8 out of 155 adjectives, 14 out of 166 nouns, and 12 out of 91 verbs were necessary. Explicit implicit negation presents a situation of a double negative, and it, therefore, reverses the negation semantics of the parts-of-speech used to indicate the presence or absence of a disease. For example, in the sentence, 'The presence of a minimal left pleural effusion cannot be excluded.', the word 'exclude' that would have indicated the absence of a disease now indicates the presence of a disease because it is negated.

Our method detected negation implicitly and also checked for explicit implicit negation. Of the 115 sentences annotated as negated in the ground-truth, an extra 33 were identified correctly through implicit and explicit implicit negation; while 93 out of the 152 unnegated sentences were identified correctly through this method. The presence of conjunctions was used to identify 17 non-negated sentences correctly, and another 41 sentences were identified correctly as unnegated because they contained the terms signifying the presence of a disease. When considering the number of sentences that could be parsed by the ACE parser, then an F1 score of 0.99 is obtained. However, when considering the entire ground-truth, including the sentences that could not be parsed, then the F1 score is 0.84.

3 Conclusion

In this paper, we have presented a method of defining a CNL from natural language by restricting the lexicon of the CNL and restricting the definition of the expression of negation. The lexicon is restricted by representing disease and anatomical medical terms as their semantic types, allowing for the processing of medical text using general-purpose semantic parsers. The second restriction of our CNL is that negation can be expressed through explicit negation, implicit negation, and explicit implicit negation. We conducted a pilot study that shows that a high F1 score (0.99) can be achieved; but also shows the limitations as a lower F1 score (0.84) results from sentences that could not be parsed. Our future work will comprise a more comprehensive evaluation of our approach, as well as seeking a solution to the problem of unparsed sentences.

References

Wendy Wibber Chapman, Will Bridewell, Paul Hanbury, F. Gregory Cooper, and G. Bruce Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.

Ann A Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage english grammar using hpsg. In *LREC*, pages 591–600. Athens, Greece.

Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):1–7.

Yang Huang and Henry J Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American medical informatics association*, 14(3):304–311.

A Johnson, T Pollard, R Mark, S Berkowitz, and S Horng. 2019. Mimic-cxr database.

Tobias Kuhn. 2014. A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170.

Woodley Packard. 2013. Ace, the answer constraint engine. *URL http://sweaglesw. org/linguistics/ace.*

Luke T Slater, William Bradlow, Dino FA Motti, Robert Hoehndorf, Simon Ball, and Georgios V Gkoutos. 2021. A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. *Computers in biology and medicine*, 130:104216.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: A fast, unsupervised approach for medical concept extraction. In *Medical Information Retrieval (MedIR) Workshop, SIGIR*, pages 1–4, Pisa, Italy.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.