Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Efficient Clinical Concept Extraction in Electronic Medical Records

Yufan Guo, Deepika Kakrania, Tyler Baldwin, Tanveer Syeda-Mahmood
IBM Research - Almaden
650 Harry Road
San Jose, California 95120

Abstract

Automatic identification of clinical concepts in electronic medical records (EMR) is useful not only in forming a complete longitudinal health record of patients, but also in recovering missing codes for billing, reducing costs, finding more accurate clinical cohorts for clinical trials, and enabling better clinical decision support. Existing systems for clinical concept extraction are mostly knowledge-driven, relying on exact match retrieval from original or lemmatized reports, and very few of them are scaled up to handle large volumes of complex, diverse data. In this demonstration we will showcase a new system for real-time detection of clinical concepts in EMR. The system features a large vocabulary of over 5.6 million concepts. It achieves high precision and recall, with good tolerance to typos through the use of a novel prefix indexing and subsequence matching algorithm, along with a recursive negation detector based on efficient, deep parsing. Our system has been tested on over 12.9 million reports of more than 200 different types, collected from 800,000+ patients. A comparison with the state of the art shows that it outperforms previous systems in addition to being the first system to scale to such large collections.

Introduction

It is well known that the structured information in electronic method record (EMR) systems does not capture all of the symptoms, diagnoses, medications, or measurements recorded in clinical reports. Automatic identification of these concepts is therefore useful not only in forming a complete longitudinal health record of patients, but also in recovering missing codes for billing, reducing costs, finding more accurate clinical cohorts for clinical trials, and enabling better clinical decision support.

The popular approach to clinical concept identification is knowledge driven, built upon various health and biomedical vocabularies such as the well-known Unified Medical Language System (UMLS) Metathesaurus. The performance of a vocabulary-based system depends on three main factors: (a) the size of the vocabulary, (b) the algorithm used for finding a match (including reliable detection of negations), and (c) its scalability in the presence of large documents and vocabularies. While a number of systems have been developed for UMLS concept identification (Aronson and Lang 2010; Savova et al. 2010; Divita et al. 2014), existing systems mostly rely on exact match retrieval from original or lemmatized reports, and very few of them are scaled up to handle large volumes of complex, diverse data. The importance of system scalability cannot be overstated, due to the exponential growth of patient data, fueling the big data revolution in healthcare.

In this demonstration we will showcase a new system for real-time detection of clinical concepts in EMR (Figure 1). The system features a large vocabulary of over 5.6 million concepts. It achieves high precision and recall, with good tolerance to typos through the use of a novel prefix indexing and subsequence matching algorithm, along with a recursive negation detector based on efficient, deep parsing. For instance, in contrast to existing algorithms, this more flexible search algorithm could detect the concept *Chest CT* within the phrase *"a subsequent CT scan of the chest"*. Similarly, the deep parsing-based negation detector captures long distance negations, such as for *Fibromyalgia* in the input *"There is no evidence suggesting the patient had been diagnosed with fibromyalgia"*. Our system has been tested on over 12.9 million reports of more than 200 different types collected from 800,000+ patients. A comparison with the state of the art shows that it outperforms previous systems in addition to being the first system to scale to such large collections.

System highlights

In contrast to existing systems, the demonstrated system proposes to show advances in both concept matching and negation detection algorithms.

Longest common word sequence matching and prefix indexing for concept extraction

We developed a novel algorithm for inexact matching of clinical terms in reports which aims to identify the longest common word sequence between a term and a sentence, an extension of the traditional longest common subsequence (LCS) problem (Maier 1978). The word sequence can be non-contiguous, but has to cover a sufficient number of words in a term, and a word is considered as a match for another if they share a prefix of sufficient size, tunable with respect to their length. Prefixes are then used for matching

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

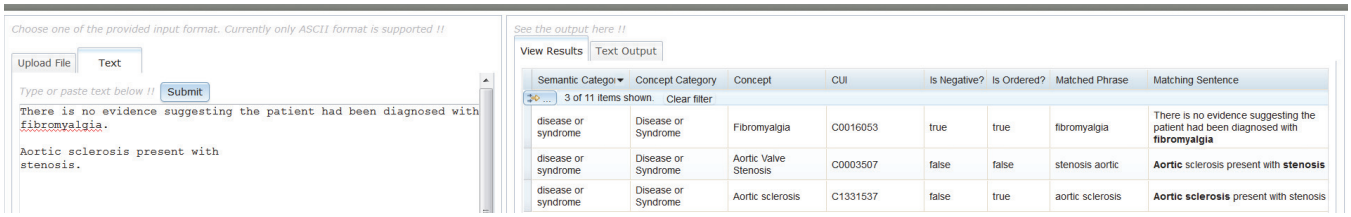

Figure 1: A snapshot of the system (featuring concept extraction, negation detection, and semantic filtering).

in correspondence to the inflection rules in English. For concepts with $\leq$3 words, order-insensitive matching is also supported.

The longest common word sequence problem is solvable in polynomial time by dynamic programming. However, finding matches for millions of terms in large collections of reports is still a computationally challenging problem. To speed up the processing, we propose a novel indexing method that significantly reduces the search space while still maintaining the requisite flexibility in matching. First, each word in the vocabulary is represented by a unique prefix, the shortest sequence of letters that differentiates it from every other term. Next, an inverted index is created for the mapping from prefixes to report sentences. Starting from the representative prefix of each term in the vocabulary (or a set of prefixes in the case of a multi-word term), all relevant sentences can be easily retrieved as potential matches for the term, and post-filtering by longest common word sequence matching can be used to further refine the search results.

On two benchmark datasets (one from the i2b2 2010 concept annotation challenge; the other from a partner hospital consisting of 700+ echocardiogram reports with disease name annotation), our concept extractor outperforms the widely-used cTAKES system (Savova et al. 2010) with significantly higher precision and recall:

	cTAKES		Our System	
	Precision	Recall	Precision	Recall
i2b2	46.7	79.6	72.6	90.4
Echocardiogram	34.0	59.6	78.2	79.3

Deep parsing-based negation detection

While many negation detectors are regular expression-based, such as the widely used NegEx algorithm (Chapman et al. 2001), there have been a few recent attempts to use sophisticated syntactic analysis for improved negation detection on linguistically complex sentences (Sohn, Wu, and Chute 2012; Mehrabi et al. 2015).

Our negation detector, built on top of the Watson deep parser (McCord, Murdock, and Boguraev 2012), differs from recent advanced negation detectors in two aspects. First, it does not require a targeted concept as input (Sohn, Wu, and Chute 2012), but returns the scope of negation in a sentence in one shot, making it possible to parallelize concept extraction and negation detection on big data. Second, it is able to capture long-distance negations within a dependency parse tree by recursively identifying negated words until the detected scope of negation converges, without limiting the "diameter" of the scope (Mehrabi et al. 2015). The recursive detection relies on a rich list of negation cues, along with carefully curated rules covering the variety of dependencies.

On a new benchmark deliberately targeted at linguistically complex sentences (1061 instances with 47.9% negated), our algorithm outperforms NegEx significantly, achieving 97% precision and 88% recall (32% and 14% higher than NegEx).

Conclusion

This paper presents a system for efficient clinical concept extraction that outperforms the state of the art through the use of novel prefix indexing, longest common word sequence matching, and deep parsing-based negation detection. In the future, the system could be linked to other knowledge bases or serve as the cornerstone of downstream applications in the clinical domain, such as question answering, information extraction, or summarization.

References

Aronson, A. R., and Lang, F.-M. 2010. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc* 17(3):229–236.

Chapman, W. W.; Bridewell, W.; Hanbury, P.; Cooper, G. F.; and Buchanan, B. G. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 34(5):301–310.

Divita, G.; Zeng, Q. T.; Gundlapalli, A. V.; Duvall, S.; Nebeker, J.; and Samore, M. H. 2014. Sophia: a expedient umls concept extraction annotator. In *AMIA Annu Symp Proc*, 467.

Maier, D. 1978. The complexity of some problems on subsequences and supersequences. *JACM* 25(2):322–336.

McCord, M. C.; Murdock, J. W.; and Boguraev, B. K. 2012. Deep parsing in watson. *IBM Journal of Research and Development* 56(3.4):3–1.

Mehrabi, S.; Krishnan, A.; Sohn, S.; Roch, A. M.; Schmidt, H.; Kesterson, J.; Beesley, C.; Dexter, P.; Schmidt, C. M.; Liu, H.; et al. 2015. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *J Biomed Inform* 54:213–219.

Savova, G. K.; Masanz, J. J.; Ogren, P. V.; Zheng, J.; Sohn, S.; Kipper-Schuler, K. C.; and Chute, C. G. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J Am Med Inform Assoc* 17(5):507–513.

Sohn, S.; Wu, S.; and Chute, C. G. 2012. Dependency parser-based negation detection in clinical narratives. In *AMIA Jt Summits Transl Sci Proc*, 1–8.