Journal of Biomedical Informatics 54 (2015) 1–9

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Domain adaption of parsing for operative notes

Yan Wang [a], Serguei Pakhomov [a,b], James O. Ryan [a,1], Genevieve B. Melton [a,c,*]

[a] Institute for Health Informatics, University of Minnesota, Minneapolis, MN, United States
[b] College of Pharmacy, University of Minnesota, Minneapolis, MN, United States
[c] Department of Surgery, University of Minnesota, Minneapolis, MN, United States

ARTICLE INFO

Article history:
Received 20 March 2014
Accepted 29 January 2015
Available online 7 February 2015

Keywords:
Probabilistic context-free grammar (PCFG)
Unlexicalized parser
Parser adaption
Natural language processing
Operative reports
SPECIALIST

ABSTRACT

Background: Full syntactic parsing of clinical text as a part of clinical natural language processing (NLP) is critical for a wide range of applications. Several robust syntactic parsers are publicly available to produce linguistic representations for sentences. However, these existing parsers are mostly trained on general English text and may require adaptation for optimal performance on clinical text. Our objective was to adapt an existing general English parser for the clinical text of operative reports via lexicon augmentation, statistics adjusting, and grammar rules modification based on operative reports.

Method: The Stanford unlexicalized probabilistic context-free grammar (PCFG) parser lexicon was expanded with SPECIALIST lexicon along with statistics collected from a limited set of operative notes tagged by two POS taggers (GENIA tagger and MedPost). The most frequently occurring verb entries of the SPECIALIST lexicon were adjusted based on manual review of verb usage in operative notes. Stanford parser grammar production rules were also modified based on linguistic features of operative reports. An analogous approach was then applied to the GENIA corpus to test the generalizability of this approach to biologic text.

Results: The new unlexicalized PCFG parser extended with the extra lexicon from SPECIALIST along with accurate statistics collected from an operative note corpus tagged with GENIA POS tagger improved the F-score by 2.26% from 87.64% to 89.90%. There was a progressive improvement with the addition of multiple approaches. Lexicon augmentation combined with statistics from the operative notes corpus provided the greatest improvement of parser performance. Application of this approach on the GENIA corpus increased the F-score by 3.81% with a simple new grammar and addition of the GENIA corpus lexicon.

Conclusion: Using statistics collected from clinical text tagged with POS taggers along with proper modification of grammars and lexicons of an unlexicalized PCFG parser may improve parsing performance of existing parsers on specialized clinical text.

1. Introduction

In the clinical domain, the rapid proliferation of patient documents within electronic health record (EHR) systems and the need to utilize these documents for secondary purposes such as disease surveillance, population health assessment, clinical research, and quality measurement have made automated information extraction and other natural language processing (NLP) techniques increasingly important. A large amount of detailed information in EHRs is stored in narrative documents, which are not directly accessible to computerized applications without specialized clinical NLP and text mining tools. NLP research to process clinical text effectively aims to improve these techniques for the specific intricacies of clinical documents.

Full syntactic parsing is an important formative step towards automated natural language understanding. Full syntactic parsing of texts provides deep linguistic features such as predicate-argument structure, voice, phrasal categories, position, and path. Moreover, incorporation of full syntactic parsing into information extraction systems has been shown to improve their performance [1–7]. Over the past decade, parsing systems have improved dramatically. Several robust parsers such as Charniak/Johnson's parser [8] and Stanford unlexicalized probabilistic context-free grammar (PCFG) parser [9] are available to produce linguistic representations for narrative text. Most of these modern parsers rely on large corpora and tag sets from general English such as

* Corresponding author at: Department of Surgery, Core Faculty, Institute for Health Informatics, University of Minnesota, 420 Delaware St SE, MMC 450, Minneapolis, MN 55455, United States. Fax: +1 612 625 4406.
E-mail address: gmelton@umn.edu (G.B. Melton).
[1] Present address: Department of Computer Science, University of California Santa Cruz, Santa Cruz, CA, United States.

http://dx.doi.org/10.1016/j.jbi.2015.01.016
1532-0464/

the Penn Treebank [10] to obtain a grammar with reasonable coverage and to acquire an accurate estimation of an appropriate statistical parsing model.

While they perform well on general English texts [12–18], these parsers may require special development and adaptation for clinical text because clinical sublanguage often differs from general English [11]. For instance, specialized domain terms and syntactic structures not typically found in general English are prevalent in clinical texts. Also, clinicians who create clinical notes have limited time and therefore frequently omit information that can be inferred from context.

Since manually annotating large numbers of parse trees is costly and may not be practical for fully supervised training within a new domain or subdomain, parser adaption is one approach proposed by researchers to improve parser performance for a domain of interest. Various methodologies have been proposed for parser domain adaption, which fall broadly into three categories: supervised domain adaption [19–21], semi-supervised domain adaption [22] and unsupervised domain adaption [12–15,17,23–25]. In supervised domain adaptation, a limited amount of labeled data from the new target domain is used to adapt the models trained on larger out-of-domain datasets. In the semi-supervised setting, the goal is to use a small amount of labeled target domain data together with lots of unlabeled data for domain adaption. In contrast, unsupervised domain adaptation relies on only unlabeled data, which is usually easy to acquire from the target domain. In principle, using a combination of limited labeled source data together with the unlabeled target data should be an effective and less costly approach to adapt an existing general English parser to the target domain.

Over the last decade, a number of techniques have been proposed for parser adaption without large amounts of manually labeled target text. Self-training is a process of taking unlabeled target text and parsing with an existing parser and add these parses to the training corpus to create a new parsing model. For example McClosky [17,26] has demonstrated that the performance of the Charniak/Johnson lexicalized PCFG parser on a target domain can be improved by including extra target domain data labeled by existing parser from the Brown corpus [26] and Medline [17]. Lexicon augmentation is another frequently used technique for parser adaption by adding extra lexical items from domain sources (e.g., Unified Medical Language System (UMLS) SPECIALIST lexicon [27]) into the existing parser lexicon. Several efforts have been devoted to improve parsing performance by extending the lexicons of parsers such as Stanford PCFG parser, Link Grammar parser, and Combinatory Categorical Grammar parser [13,14,24,25]. Finally, full parsing based on a part-of-speech (POS) tagger adapted to the target domain is also proved to be helpful for domain adaption [12,14]. A POS tagger retrained on the target domain, which is usually less expensive than retraining a parser, can provide more accurate POS tags for the back-end parsing process.

2. Background

2.1. Unlexicalized parsing and lexicalized parsing

Full syntactic parsing results in a hierarchical tree-like representation of the syntactic structure of a piece of text according to some formal grammar such as, for example, a constituency grammar [28]. Fig. 1 shows the constituency parse tree of the sentence: "The eye was patched with hyoscine ophthalmic drops."

As shown in Fig. 1, the tree representation of the input sentence from a parser conveys useful information such as the constituent boundaries, the grammatical relationship between constituents, which is expressed by the path from one constituent to another,

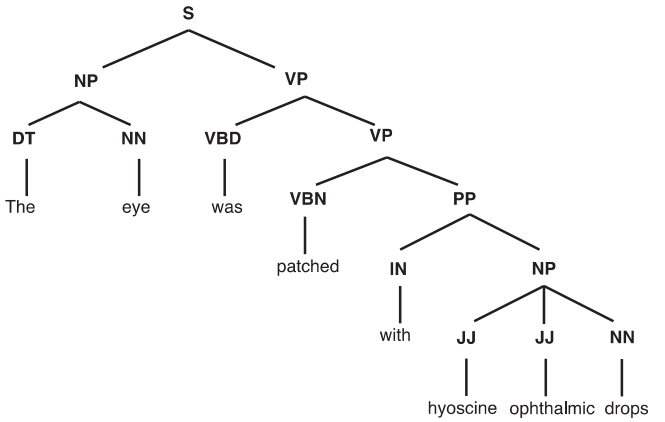

Fig. 1. Constituent (phrase structure) tree for the sentence: "The eye was patched with hyoscine ophthalmic drops." ∗S: Sentence; NP: Noun phrase; VP: Verb phrase; DT: Determiner; NN: Noun, singular or mass; VBD: Verb, past tense; IN: Preposition or subordinating conjunction; JJ: Adjective; VP: Verb phrase.

the head word of each candidate constituent and a number of other features.

In formal linguistics, Context Free Grammars [29] (CFG) are formal systems used to model natural language. CFGs contain a set of production rules (or recursive rewrite rules) that are used to generate linguistic expressions from underlying constituent building blocks. Formally, a CFG is represented as a 4-tuple consisting of 4 sets: G = (N, Σ, R, S) where:

N is a finite set of non-terminal symbols.
Σ is a finite set of terminal symbols.
R is a finite set of rules of the form $X \rightarrow Y_1 Y_2 \ldots Y_n$, where $X \in N, n \geqslant 0$, and $Y_i \in (N \cup \Sigma)$ for $i = 1 \ldots n$.
$S \in N$ is a distinguished start symbol.

For an input sequence of words, a parse tree can be derived according to the CFG production rules. Fig. 2 exemplifies a set of simple production rules. For an input sentence '*The patient left the OR*', a parse tree can be derived from the production rules as shown below in Fig. 3.

When dealing with complex natural language text, more than one production rule may apply to a sequence of words, which results in syntactic ambiguity. Fig. 3 shows two syntactic trees derived for the same sentence "*The I&A removed the viscoelastic with a tip*. . ..".

The sentences in Fig. 3 illustrate the classic phenomenon of prepositional attachment ambiguity where the interpretation of the sentence depends on whether the prepositional phrase "with a tip" attaches to the verb phrase node "removed . . ." or the lower noun phrase node "the viscoelastic."

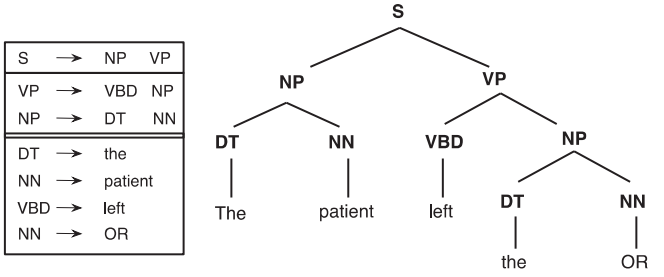

Fig. 2. Production rules example. ∗S: Sentence; NP: Noun phrase; VP: Verb phrase; DT: Determiner; NN: Noun, singular or mass; VBD: Verb, past tense; OR = Operating room.