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,

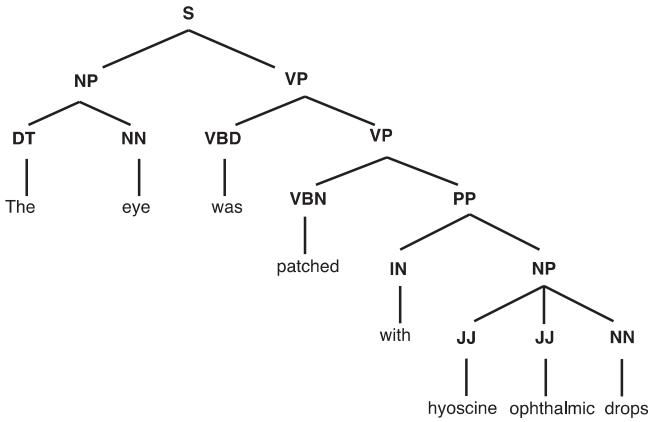

Fig. 1. Constituent (phrase structure) tree for the sentence: "The eye was patched with hyoscine ophthalmic drops." ∗S: Sentence; NP: Noun phrase; VP: Verb phrase; DT: Determiner; NN: Noun, singular or mass; VBD: Verb, past tense; IN: Preposition or subordinating conjunction; JJ: Adjective; VP: Verb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

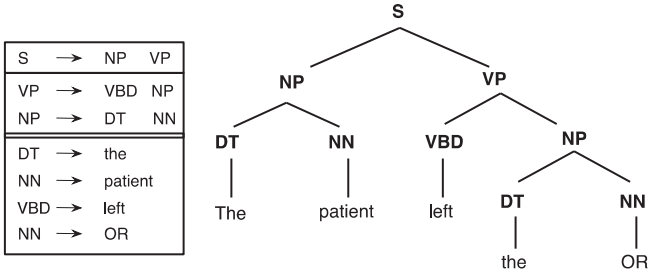

Fig. 2. Production rules example. ∗S: Sentence; NP: Noun phrase; VP: Verb phrase; DT: Determiner; NN: Noun, singular or mass; VBD: Verb, past tense; OR = Operating room.

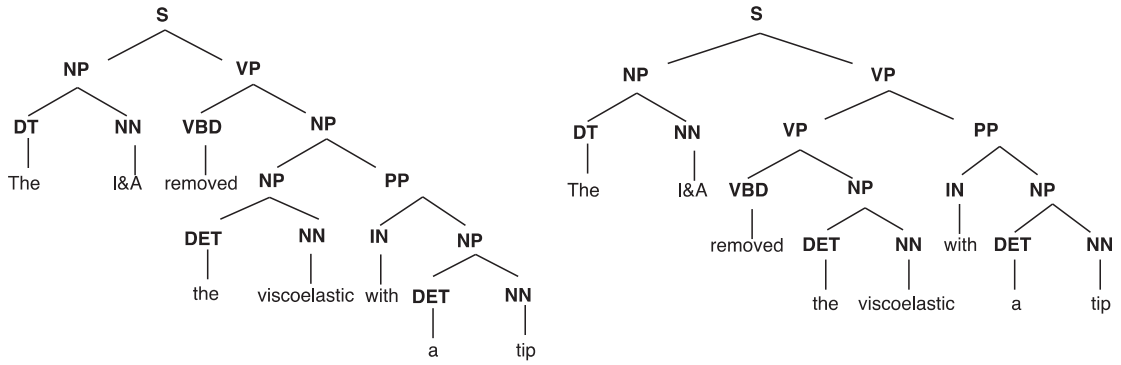

Fig. 3. Two syntactic trees for the sentence: *'The I&A removed the viscoelastic with a tip.'* ∗I&A = Irrigation and aspiration.

Probabilistic context-free grammars (PCFGs) are an attempt to deal with this ambiguity encountered when applying CFG production rules on complex natural language text. Thus, PCFG is a probabilistic version of CFG where each production has a probability, as shown in Fig. 4. In PCFG, the probability of a parse tree is the product of the probabilities of its re-write rules productions. The parse tree with the greatest probability will be picked from a number of alternatives with varying likelihoods. Probabilities of a PCFG model are typically estimated from a set of training texts (e.g., Penn Treebank [10]). Formally, a PCFG is defined as follows:

1. A context-free grammar G = (N, Σ, R, S)
2. Parameters $q(\alpha \rightarrow \beta)$, which is the conditional probability of choosing rule $\alpha \rightarrow \beta$

Given a PCFG with all parameters estimated from a corpus such as the Penn Treebank, a parse tree for a sentence s is chosen from all possible alternative parse tress by finding the parse tree with maximum likelihood:

$$\arg\max_{t \in T(s)} p(t)$$

Here t is a parse tree for s; $T(s)$ is a set of all possible parse trees for sentence s; $p(t)$ is the probability of parse tree t calculated based on parameters collected from corpus. Out-of-the-box and unenhanced PCFGs usually do not perform optimally on text from new domains [30]. Unlexicalized PCFGs with special linguistic annotations [9] and lexicalized PCFGs are two approaches that have been used to address the weaknesses of basic PCFGs.

Klein and Manning utilized a set of linguistic annotations to construct an unlexicalized PCFG parser using the probabilities associated with different syntactic categories to include vertical and horizontal history of tree nodes [9]. For example, the UNARY-INTERNAL annotation was used to mark any nonterminal node in Penn Treebank with only one child. Similarly, the TAG-PA annotation is used to mark all preterminals with their parent category as shown in Fig. 5. As shown in Klein's work, the TAG-PA annotation significantly improves parsing accuracy [9]. Here, the unlexicalized Stanford PCFG parser was trained on the Penn Treebank corpus and enriched with additional annotations and achieved similar performance to the start-of-the-art lexicalized PCFG parser without relying heavily upon lexical dependencies.

The lexicon of an unlexicalized PCFGs parser trained on treebanks with the additional annotations, as a result, stores not only lexical entries, but also the statistics that a lexical is associated with an POS tag as well as the parent tag such as "NN^NP" – a noun with a noun phrase as parent and "VBN^ADJP" – a past participle verb with an adjective phrase as parent. The grammars of an unlexicalized PCFG parser also incorporate these additional annotations. For example, a unary rule "NP^S-U -> PRN^NP" that specifies that the node has only one child. One advantage of using the unlexicalized Stanford parser is that the text format of the lexicon and grammar can be easily extended and reloaded into original parser.

A lexicalized PCFG specializes its production rules for specific words by including their head-word in the trees as shown in Fig. 6. In this way, a lexicalized PCFG largely resolves ambiguities such as the prepositional phrase (PP) attachment problem. Additionally, Collins [31] and Charniak [32] used a discriminative re-ranking technique to obtain better parse from a list of parses

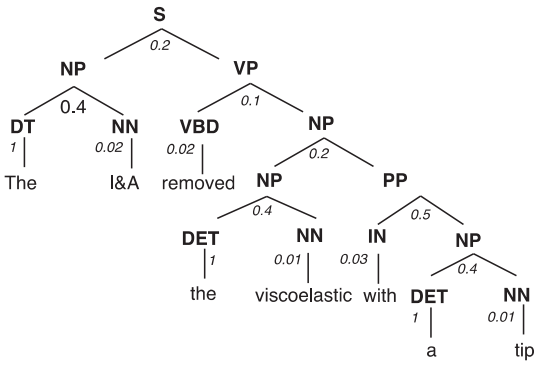

Fig. 4. A syntactic tree with production probabilities for sentence 'The I&A removed the viscoelastic with a tip.' ∗I&A = Irrigation and aspiration.

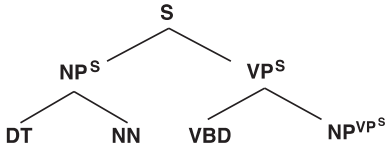

Fig. 5. Adding parent annotation to trees.

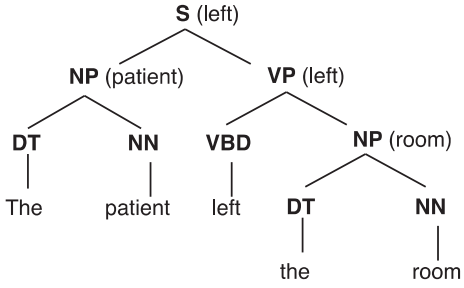

Fig. 6. Adding headtags to trees.

generated from original parsers for each sentence. However, the performance of lexicalized PCFGs is limited by the sparseness of lexical dependency information available in Penn Treebank. Also, modeling word-to-word dependencies is difficult, especially if these dependencies are domain-specific.

2.2. Domain adaption for unlexicalized parsing and lexicalized parsing

A number of groups have reported and evaluated methods to improve parsing performance of existing unlexicalized parsers. Xu and colleagues [33] reported that the use of POS tags from manual annotation could be used to produce a POS tagger for the medical domain with improved Stanford parser performance of between 2% and 4% with a small set of sentences from clinical reports. The evaluation of these enhancements revealed an improvement on the high level NLP task of noun phrase identification. Similarly, Huang et al. [23] enriched the Stanford lexicon with unambiguous entries in the SPECIALIST lexicon and customized the Stanford parser grammar based on the review of clinical reports although no formal evaluation of these modifications was performed.

We observed from preliminary experiments on clinical text particularly with operative reports that sometimes even with correct POS tags, general English parsers were not able to give correct parse tree. Fig. 7 shows parse trees of a POS tagged sentence (1) produced by the Stanford parser with and without enriched lexicons. Parse tree (7b) is produced by the original Stanford parser with correct POS tagging provided (7a) is a parse tree produced by the enriched Stanford parser.

(1) "The/DT wound/NN was/VBD extended/VBN proximally/RB and/CC distally/RB."

Self-training is a technique used to adapt a lexicalized parser to a new target domain. It creates a new parser by training an existing parser with data parsed by the existing parser as extra training data [17,26]. As shown in McClosky's work [26], the parser is retrained with the new training data set, which includes large in-domain corpus that parsed with original parser. While some early reports on self-training for parsing reported negative results, McClosky [17,26] and Bacchiani [34] have shown that this technique can improve parsing performance of the new parser on a target domain. In McClosky's work, the standard Charniak/Johnson parser was trained on a corpus of biomedical abstracts that were labeled with the existing Charniak/Johnson parser, along with Penn-Treebank. The resulted new parser showed performance improvement on a standard test set, the GENIA Treebank [35].

2.3. Procedure description section of operative reports

The narrative description of a procedure is the core portion of an operative note, which provides specific documentation of what occurred during an operation or other medical procedure (e.g., bronchoscopy, cardiac catheterization). The following text is an excerpt from a 'procedure description' section from a transurethral prostatectomy:

> *"After adequate anesthesia, the patient in the dorsal lithotomy position was prepped and draped in the usual manner. A 28 French continuous flow resectoscope sheath was inserted. Inspection showed that the patient had significant regrowth of his prostatic tissue. This patient in the past had undergone transurethral resection of the prostate elsewhere. The verumontanum and both ureteral orifices were noted to be intact. All the prostatic chips were irrigated from the bladder. A total of 46 grams of prostate was resected. Good hemostasis obtained. A 22 French three-way Foley catheter was inserted and continuous bladder irrigation was started. Sponge and needle correct X 2. The patient tolerated the procedure well."*

Effective computerized NLP systems for operative reports require an understanding of this text and ideally address the domain-specific features of operative notes. Automatic processing of such text is challenging due to higher use of certain surgery terms (e.g., "*extubate*", "*prep*"), domain-specific words (e.g., "*preperitoneally*", "*free-up*"), and special grammars (e.g., "*Good hemostasis obtained*").

2.4. GENIA corpus

GENIA corpus is a collection of articles on biological reactions of transcription factors in human blood cells. The articles are extracted from MEDLINE database with the MeSH terms, *human*, *blood cell* and *transcription factor*. Each article was annotated with parse trees following the Penn Treebank II (PTB) bracketing guidelines. The following text in Fig. 8 shows an example of GENIA syntactic annotation.

2.5. SPECIALIST lexicon

The SPECIALIST Lexicon consists of a set of lexical entries including multi-word terms with spelling variants, part(s) of speech, and other information for biomedical domain terms. SPECIALIST consists of over 200,000 biomedical terms, as well as common English words. It has been successfully used to adapt parsers for general English to the biomedical domain as it contains important syntactic, morphological, and orthographic information for each entry [13,14,23]. For instance, a lexical record for a term in

```
(S
  (NP (DT The) (NN wound))
  (VP (VBD was)
    (VP (VBN extended)
      (ADVP (RB proximally)
      (CC and)
      (ADVP (RB distally)))))
```

(a) Parse tree of sentence 1 from enriched Stanford parser

```
(S
  (NP (DT The)
    (NP
      (NP
        (NP (NN wound))
        (SBAR
          (S
            (VP (VBD was)
              (VP (VBN extended)
                (ADVP (RB proximally))))
            (NP (CC and)))))
      (ADVP (RB distally)))))
```

(b) Parse tree of sentence 1 from original Stanford parser with pre-tagged input

Fig. 7. Parse trees of a POS tagged sentence (1) produced by Stanford parser (a) with and (b) without enriched lexicon.

```
<sentence id="S2"><cons cat="S"><cons cat="NP" id="i2" role="SBJ"><cons cat="NP"><tok cat="NN">Resistance</tok>
    </cons> <cons cat="PP"><tok cat="TO">to</tok> <cons cat="NP"><tok cat="NN">glucocorticoid</tok> <tok
    cat="NN">therapy</tok></cons></cons></cons> <cons cat="VP" syn="COOD"><cons cat="VP"><tok cat="VBZ">has</
    tok> <cons cat="VP"><tok cat="VBN">been</tok> <cons cat="VP"><tok cat="VBN">observed</tok> <cons cat="NP"
    ref="i2" null="NONE"></cons><cons cat="PP"><tok cat="IN">in</tok> <cons cat="NP"><cons cat="NP"><tok
    cat="NNS">patients</tok></cons> <cons cat="PP"><tok cat="IN">with</tok> <cons cat="NP"><tok cat="JJ">
    autoimmune/inflammatory</tok> <tok cat="NNS">diseases</tok></cons></cons></cons></cons></cons></
    cons> <tok cat="CC">and</tok> <cons cat="VP"><tok cat="MD">may</tok> <cons cat="VP"><tok cat="VB">be</tok>
    <cons cat="VP"><tok cat="JJ">related</tok> <cons cat="NP" ref="i2" null="NONE"></cons><cons cat="PP"><tok
    cat="TO">to</tok> <cons cat="NP"><cons cat="NP"><tok cat="DT">the</tok> <tok cat="JJ">inflammatory</tok>
    <tok cat="NN">process</tok></cons> <cons cat="NP"><tok cat="PRP">itself</tok></cons></cons></cons></cons></
    cons></cons></cons><tok cat="PERIOD">.</tok></cons>
</sentence>
```

Fig. 8. GENIA syntactic annotation example.

Table 1
Entries of 4 POS categories in SPECIALIST lexicon and Stanford lexicon.

POS category	SPECIALIST	Stanford
Verb	56859	8477
Noun	280482	27832
Adjective	90884	9032
Adverb	12467	1422

SPECIALIST contains base forms of the term, the part-of-speech, a unified identifier, spelling variants, and inflection for nouns, verbs and adjectives. As presented in our previous work [36], the SPECIALIST lexicon has very good coverage of both verb predicates (89.9%) and nominal predicates (100%) occurring in operative notes. Table 1 shows the number of entries of four important POS categories in SPECIALIST lexicon and Stanford lexicon, demonstrating that the SPECIALIST lexicon contains many more word entries than the Stanford lexicon.

In the clinical domain, only a small amount of research has focused on parser adaption for clinical text, with previous work not focusing on operative notes. Therefore in this paper we will describe our experiments on adapting the Stanford parser for the clinical text of operative reports. We hypothesized that the addition of more accurate statistics from our clinical corpus of operative reports and use of the SPECIALST lexicon could improve the parsing performance of the Stanford parser for operative notes. We extended the lexicon of Stanford unlexicalized parser with new entries in SPECIALIST lexicon that occurred in our operative notes corpus and modified the parser grammar. We also tested the performance of parsers augmented with statistics collected from corpus POS tagged with two start-of-art POS taggers, GENIA tagger and Medpost tagger.

3. Methods

Fig. 9 provides an overview of this study. Overall, we enriched the Stanford lexicon with SPECIALIST lexicon and with statistics collected from POS-tagged operative reports from our clinical note repository and customized the Stanford grammar to the special

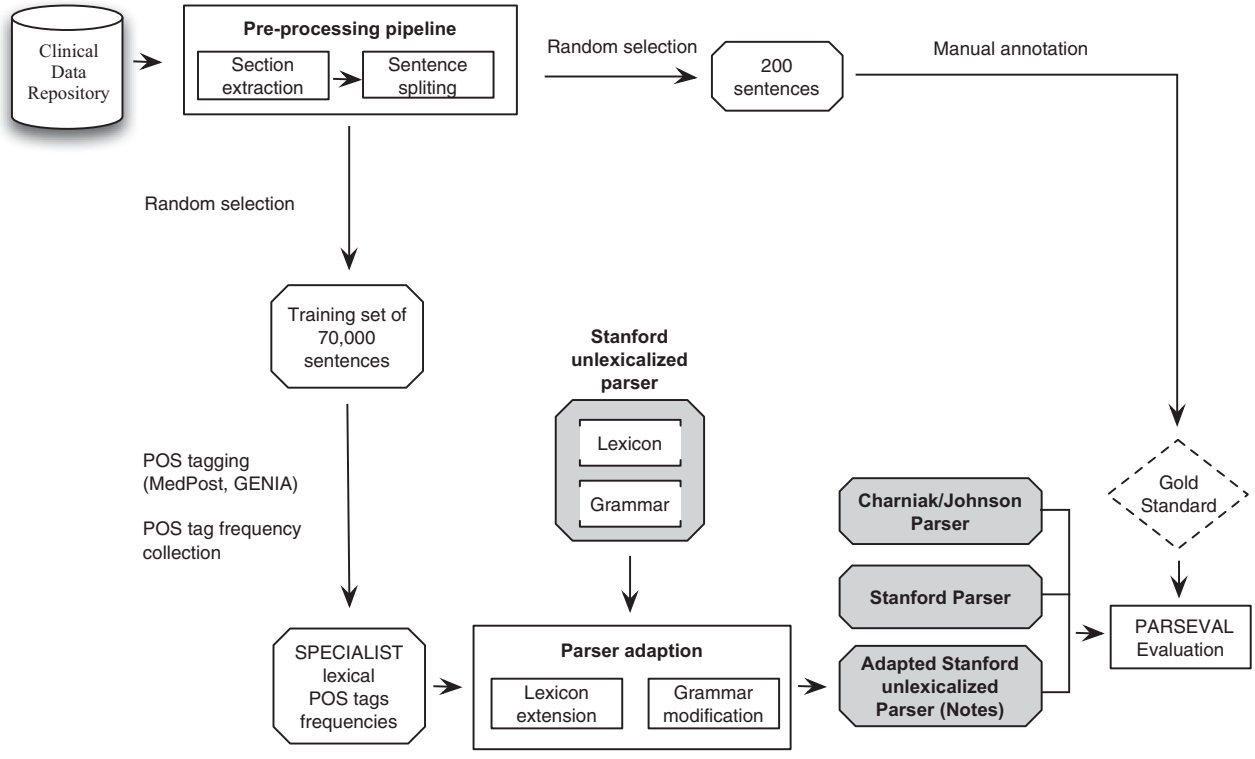

Fig. 9. Overview of operative notes parser adaption.

syntactic structure of operative report text. The resulting enhanced Stanford parser output was then evaluated and compared with POS-tagged corpus with different POS taggers using a set of manually annotated operative report sentences.

3.1. Dataset and overview

A total of 362,310 operative reports from University of Minnesota-affiliated Fairview Health Services in the Twin Cities including both community and tertiary-referral settings were used for this study. The corpus includes operative reports created by 2300 surgeons with 4333 different procedure types defined by Current Procedural Terminology (CPT) codes. The procedure description was extracted from each note and split into sentences with a locally developed heuristically-based text-processing tool (See Fig. 9, Pre-processing Pipeline). We randomly selected a dataset of 70,000 sentences, which is similar to the size of Penn Treebank, from the repository of operative notes sentences.

3.2. Stanford unlexicalized PCFG parser adaption for operative notes

The SPECIALIST Lexicon contains far more entries than the Stanford lexicon as shown previously in Table 1. To selectively expand the Stanford lexicon for operative notes, we added only SPECIALIST Lexicon entries (single word entries in this study) contained within the overall operative note corpus. This approach was taken since words that were not within the operative note corpus do not have associated frequency statistics and also to decrease the associated computational overhead encountered with loading the parser and parsing the text associated with adding a large lexicon.

In adding entries to the Stanford Lexicon, we had to take into account that the SPECIALIST Lexicon uses a set of syntactic categories that are different from the Penn Treebank tags for its entries. For unambiguous entries in the SPECIALIST lexicon, the same set of mapping rules used in Huang's work [23] were used to convert the SPECIALIST Lexicon syntactic categories into Penn Treebank tags. For ambiguous entries in the SPECIALIST lexicon, we converted those entries with multiple syntactic categories (about 20,000 words) into Stanford entries using statistics collected from the tagged corpus combined with several heuristic rules. As introduced above, an unlexicalized PCFG model requires statistics for usage of each POS tag under different parent for parsing. For instance, the word "callus" can be both a noun and a verb. To collect frequencies for the tags of each word, we first created a corpus with a similar size to the Penn Treebank from 70,000 randomly selected sentences in the operative note "procedure description" section text. Heuristic rules based on the Stanford lexicon were also used, where we observed that some parents for a particular POS tag were more frequent than others. Using adjectives as an example, in the Stanford lexicon the incidence of adjectives (68,090 in total) used within an adjective phrase (11,498) or a noun phrase (54,211) was significantly greater than other phrase types. The sentence set was then tagged using the five Stanford POS taggers. For example, in the Stanford lexicon, frequencies for each POS tag with a different parent for the word "inject" are given in Table 2. To decide the frequency distribution of each possible parent, we collected the frequency from POS tagged sentences.

We also observed that for some POS tag and parent combinations, only one or a few specific words were associated. For example, the word "only" in sentence (2) is the only adjective word that could be used in a conjunction phrase:

(2) "The biceps tendon, long head intra-articular portion, was not only split, but remarkably frayed."

Table 2
Frequency of each POS tag of word "inject" with different parents in the Stanford lexicon.

POS tag	Parent tag	Frequency
VBD	VP	2
VBN	VP	2
JJ	ADJP	7
JJ	NP	2
JJ	WHADJP	1
JJ	WHNP	1
JJ	UCP	1
JJ	QP	1

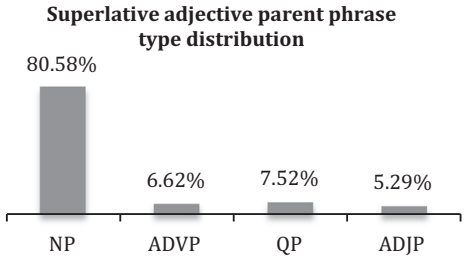

Fig. 10. Parent phrase type distribution of the POS tag superlative adjective.

Thus, for each POS tag such as "JJ", "NN" and "VBD", we defined a heuristic parent distribution for it and split the collected frequency based on these distributions. For example, for POS tag "JJS" (superlative adjective), we define a distribution as shown in Fig. 10. From each POS tagged corpus, the frequency of POS tags associated with each SPECIALIST lexicon entry within the set of 70,000 sentences was collected and used to adapt the Stanford lexicon and create a new adapted lexicon. For example, the new Stanford lexicon extended with the MedPost lexicon contained 172,636 entries while the original Stanford lexicon had 101,703 entries.

Using our previous observation that physicians tend to use passive voice to narrate the procedure description section [36], we manually adjusted the frequencies of VBD (verb, past tense) and VBN (verb, past participle) tags for verb entries that could be both a past tense verb and a past participle. Also, the POS tag of some verbs, such as "appeared", "tolerated" and "revealed", can be either VBD and VBN in the SPECIALIST lexicon, but after review of a random set of sentences with these words, we found that the POS tags of these verbs were mostly VBD as opposed to other verbs such as "incised" and "dissected" which tended to mostly be used in text as VBN. To assign frequencies that could better reflect actual usage of verbs, we used the 200 verbs previously reported that covers 92% of all verbs from operative notes to help provide reasonable frequencies of potential ambiguous POS tags.

Finally, we were able to omit auxiliary verbs as this was another feature previously observed in the sublanguage of operative notes. For example, in following sentences (3) and (4), the auxiliary verb "was" is omitted in the operative note text.

(3) "A transverse incision was made in the popliteal fossa and the lesser saphenous vein identified, ligated proximally."
(4) "Good hemostasis obtained."

Syntactical information such as the voice of verbs is also critical for many NLP tasks such as semantic role labeling. To address this problem in operative notes, we modified the grammar of the Stanford parser by including more productions rules. For example, given sentence "Good hemostasis obtained" original sentence will

give a parse tree as (5). After adding a new rule "VP^S-VBF-v -> VBN^VP", the parser gives correct parse as (6). The new parse assigns correct phrase tags and POS tags for the verbs, which are very important to NLP tasks such as semantic role labeling [37]. As shown in Gildea's work the phrase tags and POS tags are used to extract voice and parse tree path for semantic role calculating.

(5) (ROOT (S (NP (NNP Gelfoam)) (VP (VP (VBD applied)) (CC and) (ADVP (RB hemostasis)) (VP (VBD confirmed))))).
(6) (ROOT (S (NP (NP (NNP Gelfoam)) (VP (VP (VBN applied)) (CC and) (VP (NN hemostasis)))) (VP (VBN confirmed)))).

4. Evaluation

To evaluate the performance of parsers adapted from the corpus POS-tagged using different POS taggers, we created a reference standard with 200 manually annotated parse trees of randomly selected operative notes sentences. The reference standard parse trees were annotated by two separate annotators with both a linguistics and informatics background and experience in clinical NLP. Annotations followed the Penn Treebank II Bracketing guidelines [38]. To compare parse results of adapted parsers with the parse trees produced by the Charniak/Johnson parser, parse trees generated by the original Stanford parser and parse trees generated by the original Stanford parser with POS tags from MedPost were examined. In addition, we tested the performance of the parser on a random set of GENIA parse trees. Since the GENIA corpus is from a slightly different domain, we wanted to evaluate the same technique for parser adaption on this domain.

Parsing performance was evaluated following the PARSEVAL standards [39] for parsing accuracy evaluation. Each constituent in the parse was represented as a labeled span. A constituent is counted as correct only if the label and text span is correct. Given two parses, the precision and recall of constituents were calculated. Precision and recall can be formally defined in terms of the number of true positive (TP), false positive (FP) and false negatives (FN) as below. F-score is the weighted harmonic mean of precision and recall. Syntactic annotations from two annotators for the same evaluation set of a 10% sample of the full evaluation set were compared and the proportion agreement of annotations was computed at the sentence level.

To evaluate the significance of parsing performance differences between the parsers, a pair-wise Wilcoxon Signed-Rank test with Bonferroni adjustment was conducted on the F-scores of the parsers evaluated on the test set sentences. As the F-score differences between parsers severely deviated from a Gaussian distribution, the Wilcoxon Signed-Rank test was used for a statistical evaluation since it does not require a normal distribution of differences between data pairs as required by the pair-wise t-test.

To evaluate the proposed parser adaption technique, a similar approach to the parser adaption for operative notes was used to adapt Stanford unlexicalized PCFG parser for the GENIA corpus. We used 14,325 training trees from the GENIA Treebank as a training corpus and collected statistics from it. Since we did not have enough biology domain knowledge, the words that occurred in GENIA were simply ported into the Stanford unlexicalized PCFG parser lexicon. Since GENIA trees have parent labels for each word, we tested our approach with two sets of lexicons, one with the accurate parent statistics and the other one with parent statistics generated from heuristics rules. We removed old entries in the original Stanford lexicon when the entry exists in the GENIA corpus. A simple grammar was added into the Stanford lexicon for testing resulting in about 129,600 entries for the new parser.

5. Results

The inter-rater agreement between the two annotators for the syntactic tree annotation task was 85%. For most sentences, the two annotators agreed on all the phrase tags and POS tags in the syntactic tree. In the six sentences where the annotators did not agree, there were minor differences in annotations in three sentences and major differences in three sentences. The three sentences with major differences in annotations tended to be complex sentences such as the following sentence: "Following induction of general anesthesia, intubation with a bronchial blocker, positioning in the right lateral decubitus position, the left chest was prepped and draped and a total, ultimately, of 3 port incisions were made.".

The precision, recall, and f-score means for each of the parsers evaluated are summarized in Table 3 for parsers adapted for operative notes and in Table 4 for those adapted for the GENIA corpus. As shown in Table 3, at baseline, the Charniak/Johnson parser had slightly better parsing performance for operative notes compared to the Stanford parser. The expansion of the lexicon yielded moderate improvement in parsing performance. Grammar modification combined with statistics adjustment also resulted additional performance gain. The f-score of the final adapted Stanford parser on the operative notes test set improved from 87.64% to 89.90%. The pair-wise t Wilcoxon Signed-Rank Test with

Table 3
Evaluation results of parser adaption for operative notes.

Parser	Precision (%)	Recall (%)	F-score (%)
Evaluation of parser adaption for operative notes			
Baseline (stanford unlexicalized parser)	87.54	87.74	87.64
Charniak/Johnson	88.43	88.46	88.45
Adapted stanford unlexicalized parser (New grammars)	87.73	87.94	87.83
Adapted stanford unlexicalized parser (Lexicon expansion)	88.82	89.28	89.04
Adapted stanford unlexicalized parser (New grammars + lexicon expansion)	89.27	89.84	89.55
Adapted stanford unlexicalized parser (New grammars + lexicon expansion + statistics adjustment)	89.65	90.13	89.90

Table 4
Evaluation results of parser adaption for GENIA.

Parser	Precision (%)	Recall (%)	F-score (%)
Evaluation of parser adaption for GENIA corpus			
Baseline (stanford unlexicalized parser)	78.18	73.52	75.78
Adapted stanford unlexicalized parser (New lexicon with parent statistics by rules and new grammar)	82.92	76.52	79.59
Adapted stanford unlexicalized parser (New lexicon with actual parent statistics and new grammar from)	84.08	78.60	81.25

Bonferroni adjustment for F-scores of the parsers shows parsing performance improvement of the best-adapted parser to the baseline parser (Stanford unlexicalized parser) (p-value < 0.001).

Table 4 shows the performance of the parser adapted on the GENIA corpus, when apply same technique on GENIA corpus, the parsing result of adapted parser on the GENIA test set improved from 75.78% to 79.59% with parent distribution from rules and to 81.25% with parent distribution collected from GENIA Treebank annotations.

6. Discussion

Full syntactic parsing of text provides deep linguistic information (e.g. voice, phrase type) useful for many NLP tasks. Parsers developed for general English text have benefited from a large tree bank and training corpus (e.g., Penn Treebank) and have achieved high parsing performance. Clinical documents are known to have special sub-language features (e.g. domain vocabulary, telegraphic text, special grammar), which often require adaptation of general English NLP tools. Parsers often have a decrement in performance when applied to scientific texts [18]. Domain NLP experts have investigated methods to adapt parsers trained on general English to new target domains [12–15,17,18,40,41]. However, these approaches have been attempted to only a limited extent in some types of clinical texts. In this work, we investigated the adaptation of a general unlexicalized PCFG parser to a specific type of clinical text - operative reports using tag statistics collected from operative reports and other sublanguage features of operative notes. We applied the approach on two different domains, clinical operative notes and the GENIA corpus. The results show that this approach can improve parsing performance on both domains. Though an increase of 2.26% of the parsing performance on operative notes is not large in absolute performance, this improvement is still noteworthy as the baseline performance of the unlexicalized PCFG parser very was good at operative notes. As shown in our results, domain adaptation was helpful in improving parser performance further. We plan to incorporate the adapted parser into our NLP system, the biomedical information collection and understanding system (BioMedICUS [42]).

To compare our results with previously work on parser domain adaption, we applied our approach on the GENIA corpus, which is a public available corpus. Our evaluations show that the performance of the new parser adapted to GENIA corpus is close to the state of the art parser performance 80.7% without parser training using domain parse trees [43], which requires a large annotated corpus and is not feasible for parser adaption in most cases.

To extend the Stanford parser lexicon, we incorporated only the SPECIALIST entries that existed in our corpus. Another option to consider with future enhancements would be to add all tokens in the operative notes corpus, which would not limit us to the ones contained in the SPECIALIST lexicon. We observed that out of all the tokens in our corpus, about 75% of them were contained in the SPECIALIST lexicon. Some tokens in our corpus are not counted as in SPECIALIST lexicon because that the first letters of these words are capitalized since that the Stanford unlexicalized PCFG parser treat upper cased word and lower cased word differently. Of all of the tokens not in the SPECIALIST lexicon, a large portion of them (about 85%) were nouns. Since the Stanford parser treats unknown words as nouns by default, we chose to ignore these tokens. However, we did include adjective and adverb tokens, which are in our corpus but not in the SPECIALIST lexicon because of capitalization of the first letter when these words appear at the beginning of a sentence. In this study, only single words entries in SPECIALIST lexicon were incorporated into the Stanford lexicon. More research and experiments will needed to incorporate multi-word entries in future study.

In this work, we used a set of heuristic rules to specify the parent distribution of each entry depend on the POS tag of the token as shown in section 3.2. As shown in Table 4, when use real parent phrase tag distribution collected from GENIA tree bank, the adapted parser performance improved another 1.68%. However, real parent phrase tag distribution is not always available for other domain such as the clinical text. To acquire a better estimation of the statistics on parent distribution, some features such as the POS tag of the word before and after the interested word may help to decide the parent phrase tag. More work will be needed to analysis the algorithm for parent distribution in the future. When tested the new unlexicalized PCFG parser adapted with clinical text on GENIA tree bank, as we expected, we found no performance improvement. As the GENIA corpus is a domain with very different sublanguage features, the statistics of GENIA text have differences from clinical text.

Since the Stanford PCFG parser is unlexicalized, no head word information is incorporated in the associated production rules. Thus, we observed that the adapted Stanford parser was unable to solve the prepositional phrase (PP) attachment ambiguity, which an issue often observed in general English. In the text for procedure description, we observed that the average sentence length (86 characters) is less than that of the Wall Street Journal sentences (126 characters). As shown in the example procedure description in the introduction, surgeons tend to describe actions, which occurred during a procedure using short and simple sentences. Thus, the ambiguity is potentially less of a problem in operative notes than in general English and other clinical texts.

In addition, since procedures in operative notes are usually described with short and simple sentences, the parsing performance of regular parsers is better than that of some other types of clinical text such as the corpus presented in Xu's work [33]. Other areas where we might consider further study include increasing the parse tree training set, which we purposefully did not do here with the goal of enhancing the parser with corpus statistics and other sublanguage characteristics. Subjectively, the overall parsing performance improvement observed with these enhancements was good despite the small magnitude of increase observed since the baseline performance of the unenhanced Stanford parser was fairly high. Furthermore, the magnitude of increase in performance accuracy found in this study is consistent with that found in other similar studies of parser adaptation [12,14,33,40].

While the operative notes dataset is relatively small and is a limitation of the study, the dataset is unique in nature and labor intensive to create. Other publicly available labeled clinical corpora for research contain few operative notes, such as the MiPACQ [44] corpus which contains only one operative note. We also evaluated our parser adaption technique using the GENIA tree bank for biology text and observed similar results. As additional publically available tree banks are become available, it would be value to perform other parallel, independent evaluations to test if this approach is more generalizable.

In placing this study in the overall context of clinical NLP, we only concentrated on the clinical text for the procedure description of operative notes. Additional work will be needed to determine if the approach used here with operative reports will be generalizable to other types of clinical texts such as discharge summaries and radiology reports. These approaches may require a good understanding and consideration of other unique syntactic structures and language features seen in clinical documents, such as the irregular sentence structures observed in Xu's work [33]. We suspect that by including additional grammars for irregular structures into the Stanford parser and extending the parser lexicon to the lexicon specific to those texts that the performance of

the Stanford parser can similarly be improved on other clinical text in an analogous manner.

7. Conclusion

In this study, we adapted the Stanford parser by extending the parser lexicon with new entries in operative notes. Syntactical statistics of each new entry were collected from POS tagged clinical text. The 200 most frequent verbs were modified in their entries using an adjustment based on their usage in operative notes. The Stanford parser unary and binary grammar were also customized to deal with the special syntactical structure of operative notes. Our experiments showed that augmenting the lexicon combined with statistics collected from GENIA tagged corpus of operative notes and new production rules best augmented the parsing performance of Stanford parser. When applying a similar approach on the GENIA Treebank corpus, we observed similar improvement with the adapted unlexicalized parser by augmenting the lexicon and grammar production rules.

Acknowledgments

The author would like to thank Fairview Health Services and support from the American Surgical Association Foundation (GM), #R01 LM009623-01 (SP) National Library of Medicine, Institute for Health Informatics Seed Grant (GM/SP), and University of Minnesota Clinical Translational Science Award 8UL1 TR000114-02.

References

[1] Kilicoglu H, Bergler S. Syntactic dependency based heuristics for biological event extraction. In: Proceedings of the workshop on current trends in biomedical natural language processing: Shared Task; Boulder, Colorado; 2009. p. 119–27.
[2] Rinaldi F, Schneider G, Kaljurand K, Hess M, Romacker M. An environment for relation mining over richly annotated corpora: the case of GENIA. BMC Bioinformatics 2006;7(Suppl. 3):S3.
[3] Rinaldi F, Schneider G, Kaljurand K, Hess M, Andronis C, Konstandi O, et al. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. Artif Intell Med 2007 February;39(2):127–36.
[4] Song M, Yu H, Han WS. Combining active learning and semi-supervised learning techniques to extract protein interaction sentences. BMC Bioinformatics 2011;12(Suppl. 12):S4.
[5] Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics 2004;20(5):604–11. March 22.
[6] Bui Q-C, Nuallain B, Boucher C, Sloot P. Extracting causal relations on HIV drug resistance from literature. BMC Bioinformatics 2010;11(1):101.
[7] Rinaldi F, Schneider G, Clematide S. Relation mining experiments in the pharmacogenomics domain. J Biomed Inform 2012;45(5):851–61. October.
[8] Charniak E. A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference; Seattle, Washington; 2000. p. 132–39.
[9] Klein D, Manning C. Accurate unlexicalized parsing. In: Proceedings of the 41st annual meeting of the association for computational linguistics: 7–12 July 2003; Sapporo; 2003. p. 423–30.
[10] Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: the Penn treebank. Comput Linguist 1993;19(2):313–30.
[11] Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002;35(4):222–35. August.
[12] Rimell L, Clark S. Porting a lexicalized-grammar parser to the biomedical domain. J Biomed Inform 2009;42(5):852–65. October.
[13] Szolovits P. Adding a medical lexicon to an English Parser. In: AMIA annu symp proc; 2003. p. 639–43.
[14] Pyysalo S, Salakoski T, Aubin S, Nazarenko A. Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. BMC Bioinformatics 2006;7(Suppl. 3):S2.
[15] Lease M, Charniak E. Parsing biomedical literature. In: The Second international joint conference on natural language processing (IJCNLP-05); 2005. p. 58–69.
[16] Clegg A, Shepherd A. Benchmarking natural-language parsers for biological applications using dependency graphs. BMC Bioinformatics 2007;8(1):24.
[17] McClosky D, Charniak E. Self-training for biomedical parsing. In: Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: short papers; Columbus, Ohio; 2008. p. 101–4.
[18] Clegg AB, Shepherd AJ. Evaluating and Integrating Treebank Parsers on a Biomedical Corpus. In: Proceedings of the ACL workshop on software; 2005. p. 14–33.
[19] Hara T, Miyao Y, Tsujii Jun'ichi. Adapting a probabilistic disambiguation model of an HPSG Parser to a New Domain. In: Natural language processing – IJCNLP 2005. Berlin, Heidelberg: Springer; 2005. p. 199–210.
[20] Daumé III H. Frustratingly easy domain adaptation. In: Proceedings of the 45th annual meeting of the association of computational linguistics; 2007. p. 256–63.
[21] McClosky D, Charniak E, Johnson M. Effective self-training for parsing. In: HLT-NAACL '06 Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics; 2006. p. 152–59.
[22] Daumé III H, Kumar A, Saha A. Frustratingly easy semi-supervised domain adaptation. In: Proceedings of the 2010 workshop on domain adaptation for natural language processing; Uppsala, Sweden. 1870534: Association for Computational Linguistics; 2010. p. 53–9.
[23] Huang Y, Lowe HJ, Klein D, Cucina RJ. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. J Am Med Inform Assoc 2005;12(3):275–85.
[24] Swift M, Blaylock N, Allen J, Beaumont Wd, Galescu L, Jung H. Augmenting a Deep Natural Language Processing System with UMLS. In: Proceedings of the fourth international symposium on semantic mining in biomedicine (SMBM 2010); Hinxton, UK; October 2010.
[25] Oostdijk N, Verberne S, Koster CHA. Constructing a Broad-coverage Lexicon for Text Mining in the Patent Domain. In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10); Valletta, Malta; 2010.
[26] McClosky D, Charniak E, Johnson M. Reranking and self-training for parser adaptation. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics; Sydney, Australia; 2006. p. 337–44.
[27] SPECIALIST Lexicon. Available from: <http://www.nlm.nih.gov/pubs/factsheets/umlslex.html> [accessed 14.01.15].
[28] Sipser M. Introduction to the theory of computation. International Thomson Publishing; 1996.
[29] Chomsky N. Syntactic structures. Mouton de Gruyter; 2002.
[30] Collins M. Head-driven statistical models for natural language parsing. Comput Linguist 2003;29(4):589–637.
[31] Collins M, Koo T. Discriminative reranking for natural language parsing. Comput Linguist 2005;31(1):25–70.
[32] Charniak E, Johnson M. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: ACL'05 Proceedings of the 43rd annual meeting on association for computational linguistics; 2005. p. 173–80.
[33] Hua X, AbdelRahman S, Min J, Jung-wei F, Yang H. An initial study of full parsing of clinical text using the Stanford Parser. In: Proceedings of the 2011 IEEE international conference on bioinformatics and biomedicine workshops, 2011 12–15 November; 2011. p. 607–14.
[34] Bacchiani M, Riley M, Roark B, Sproat R. MAP adaptation of stochastic grammars. Comput Speech Lang 2006;20(1):41–68.
[35] Genia treebank. Available from: <http://www.nactem.ac.uk/genia/genia-corpus/treebank> [accessed 14.01.15].
[36] Wang Y, Pakhomov S, Burkart NE, Ryan JO, Melton GB. A study of actions in operative notes. AMIA Annu Symp Proc, 2012; 2012. p. 1431–40.
[37] Gildea D, Palmer M. The necessity of parsing for predicate argument recognition. In: Proceedings of the 40th annual meeting on association for computational linguistics; 2002. p. 239–46.
[38] Penn Treebank II Bracketing Guide; 1995. Available from: <ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz> [accessed 14.01.15].
[39] Abney S, Flickenger S, Gdaniec C, Grishman C, Harrison P, Hindle D, et al. Procedure for quantitatively comparing the syntactic coverage of English grammars. In: Proceedings of the workshop on Speech and Natural Language; 1991. p. 306–11.
[40] Rush AM, Reichart R, Collins M, Globerson A. Improved parsing and POS tagging using inter-sentence consistency constraints. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning; 2012. p. 1434–44.
[41] Miller JE, Torii M, Vijay-Shanker K, editors. Subdomain adaptation of a POS tagger with a small corpus. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; 2006.
[42] BioMedICUS. Available from: <https://bitbucket.org/nlpie/biomedicus> [accessed 14.01.15].
[43] McClosky D. Any domain parsing: automatic domain adaptation for natural language parsing. Brown University; 2010.
[44] Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ clinical question answering system. AMIA Annu Symp Proc, 2011; 2011. p. 171–80.