Improved Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon

YANG HUANG, MS, HENRY J. LOWE, MD, DAN KLEIN, PHD, RUSSELL J. CUCINA, MD, MS

Abstract

Objective: The aim of this study was to develop and evaluate a method of extracting noun phrases with full phrase structures from a set of clinical radiology reports using natural language processing (NLP) and to investigate the effects of using the UMLS® Specialist Lexicon to improve noun phrase identification within clinical radiology documents.

Design: The noun phrase identification (NPI) module is composed of a sentence boundary detector, a statistical natural language parser trained on a nonmedical domain, and a noun phrase (NP) tagger. The NPI module processed a set of 100 XML-represented clinical radiology reports in Health Level 7 (HL7)® Clinical Document Architecture (CDA)–compatible format. Computed output was compared with manual markups made by four physicians and one author for maximal (longest) NP and those made by one author for base (simple) NP, respectively. An extended lexicon of biomedical terms was created from the UMLS Specialist Lexicon and used to improve NPI performance.

Results: The test set was 50 randomly selected reports. The sentence boundary detector achieved 99.0% precision and 98.6% recall. The overall maximal NPI precision and recall were 78.9% and 81.5% before using the UMLS Specialist Lexicon and 82.1% and 84.6% after. The overall base NPI precision and recall were 88.2% and 86.8% before using the UMLS Specialist Lexicon and 93.1% and 92.6% after, reducing false-positives by 31.1% and false-negatives by 34.3%.

Conclusion: The sentence boundary detector performs excellently. After the adaptation using the UMLS Specialist Lexicon, the statistical parser’s NPI performance on radiology reports increased to levels comparable to the parser’s native performance in its newswire training domain and to that reported by other researchers in the general nonmedical domain.


The medical record in the United States is still largely paper based. However, there is increasing interest in the creation of a national model for a ubiquitous electronic health record (EHR).1 To ensure wide adoption and interoperability, the EHR must be based on standards. In particular, the use of standard terminologies for data representation will be critical. Many clinical information systems enforce standard semantics by mandating structured data entry. While this approach can be successfully applied for a subset of core clinical data elements, it has limited use when dealing with the myriad of narrative clinical documents that comprise the majority of the data making up the patient record. These documents are often dictated and then transcribed into electronic format, with little or no attempt to standardize content representation.

The Health Level 7 (HL7)® Clinical Document Architecture (CDA)2 offers a standard for the representation and communication of clinical documents but currently leaves the methodology for representing document content to the system implementer. We have been interested in this problem for some time, in part because of the importance of automatically linking imaging data to clinical imaging reports using standardized terminology in Multimedia Electronic Medical Record Systems (MEMRS).3 We are developing a system called ChartIndex that transforms electronic clinical documents into an XML-based, CDA-compliant format and then automatically identifies and represents important biomedical concepts within the transformed documents using the National Library of Medicine’s (NLM) Unified Medical Language System (UMLS).4,5

Many researchers have worked on the problem of automated biomedical concept recognition. The SAPHIRE system designed by Hersh et al.6,7 automatically encodes UMLS concepts using lexical mapping. The lexical approach is computationally fast and useful for real-time applications.
However, this approach when used alone may not provide optimal results. More recently, Zou et al.\(^8\) developed IndexFinder to add syntactic and semantic filtering to improve performance on top of lexical mapping. Other researchers use more advanced Natural Language Processing (NLP) techniques such as part of speech (POS) tagging and phrase identification together with lexical techniques to facilitate concept indexing either in clinical documents\(^9,10,11,12,13\) or biomedical literature retrieval.\(^14,15,16,17,18,19,20\) MedLEE\(^11\) is a system developed by Friedman et al. to encode free clinical text into structured format. Along with a few other systems,\(^21,22,23\) MedLEE encodes modifiers together with core concepts in noun phrases (NPs). It has been applied to a number of different types of clinical documents and achieved encouraging results.\(^24,25,26\) Cooper and Miller\(^12\) carried out an evaluation on MeSH\(^\text{®}\) encoding using a lexical approach, statistical approach, and a hybrid approach. Nadkarni et al.\(^13\) conducted a study of concept matching in discharge summaries and surgical notes. Purcell and Shortliffe\(^15\) used concepts encoded in document headings to improve indexing of free text. Berrios\(^16\) exploited the above technique together with a vector space model and a statistical method to match text content to a set of query types. Berrios et al.\(^27\) also reported on building a semi-automatic system called Internet-based Semi-Automated Indexing of Documents (ISAID) to aid the manual indexing of textbook. MetaMap\(^7,18\) identifies UMLS concepts in text and returns them in a ranked list in a five-step process, identifying simple NPs, generating variants of each phrase, finding matched phrases, assigning scores to matched phrases by comparing them with the input and composing mappings. MetaMap has been used in a number of applications.\(^28,29\)

In almost all of the above systems, phrase identification is an important step. Phrase identification, especially noun phrase identification (NPI), has been investigated by researchers in both general domains and the biomedical domain. In general domains, Bourigault\(^30\) reported using maximal NPs to extract terminologic NPs. However, no performance value was reported. Ramshaw and Marcus\(^31\) and Cardie and Pierce\(^32\) reported precision and recall of 91% to 93.5% on extracting base NPs from the Penn Treebank Wall Street Journal (WSJ) corpus. As mentioned by Cardie and Pierce, the above work is difficult to compare directly because each approach used a slightly different definition of NPs (even for base NPs).

In the biomedical domain, Bennett et al.\(^33\) reported the precision and recall of using several NPI systems on Medline abstracts. The performance was measured on four fields: title, author, abstract, and MeSH\(^\text{®}\) terms and may not be comparable to NPI on free text alone, as only the abstract field in the above contains free text in complete sentences, whereas the text in the other three fields are most likely already in NPs. The above two systems were not augmented with the UMLS Specialist Lexicon.\(^34\) Spackman and Hersh\(^35\) reported NPI performance on discharge summaries. The recall and precision on NPs were 77% and 68.7%, respectively, using two different systems. The results were better if measured on partial NPs. Berrios et al.\(^36\) reported in their experiment that 66% of nouns and noun modifiers and 81% of NPs were correctly matched to UMLS concepts. However, the experiment focused on the evaluation of a concept matching algorithm, and no comparable performance numbers on NPI were reported.

Our prior work\(^37,38\) has been affected by poor precision. We have previously described\(^39\) a new approach, called contextual indexing, that can partially ameliorate this problem. However, this approach is dependent on implementing a successful NPI engine. In this report we describe a successful strategy, based on machine learning, statistical natural language processing, and the use of the UMLS Specialist Lexicon, which is successful at both sentence boundary detection and biomedical NPI within clinical radiology reports. Rather than develop a system that results in a single NP representation, we have instead adopted a deep parsing approach that captures the entire parse tree for each target NP within a clinical document. We believe that this approach offers maximum flexibility at the time of indexing. To evaluate this approach we performed two experiments, described in this report, which used the maximal NP and the base NP within each tree to calculate boundary precision and recall characteristics of the model.

**Methods**

Our approach to encoding biomedical concepts in clinical documents makes the following assumptions: most biomedical concepts are represented as NPs; most biomedical concept NPs do not span across sentences; NPs serve as a good starting point for UMLS indexing. With the above assumptions, we set out to develop and evaluate robust automated mechanisms for delimiting sentences and identifying NPs in clinical radiology reports.

**Setting and Selection of Documents**

The document collection used in this study was 100 radiology reports, randomly selected from a larger collection of 1,000 de-identified radiology reports from Stanford Hospital & Clinics. The reports covered the most common imaging modalities (25 computed tomography [CT] scans, five mammograms [MAMO], 25 magnetic resonance imaging [MR/MRI] reports, 10 radiology procedures [PROC], 30 radiographs [RAD], and five ultrasounds [US]). There were 16,298 words, 3,043 maximal NPs, 4,755 base NPs and 1,506 sentences in this document set. They were mostly well-formed sentences, with partial sentences almost all in NPs. This set of 100 documents was split into two halves serving as a training set and a test set.

**Data Collection and Processing**

As reported in our prior work,\(^39\) we have developed a software module to reliably convert semistructured free-text radiology reports into segmented HL7 CDA-compatible XML documents.\(^2\) In this experiment, using these XML documents we implemented a three-step process to identify NPs: sentence boundary detection (SBD), full NLP parsing, and NPI tagging.

**Sentence Boundary Detection**

While the parser we used did provide a crude sentence-breaking mechanism, its accuracy was unsatisfactorily low. Therefore, we developed a sentence boundary detector to pre-delimit sentences within documents before sending them to the parser. Among many potential machine-learning methods for SBD, such as Naive Bayes, Decision Tree, Neural Network, Hidden Markov Model (HMM), and Maximum Entropy Modeling (MEM), we decided to choose MEM as our main method of solving this problem. This decision was based on a number of factors: first, the MEM has a solid
information and bi-grams can be modeled easily as features in an integrated model. Also, features no longer have to be limited to local context. For example, the following text:

"his compromise bill." A committee staffer

is highly ambiguous. We cannot tell from the above local context, if the period or the double quote should be the sentence boundary. However, in MEM, we can easily add the parity of double quotes (i.e., if the quotes are opening or closing quotes) as a feature, which is highly effective, but cannot be derived from local context in the above example.

MEM automatically calculates weights of features in training and handles overlapping features very well. Thus, in MEM implementations, it is sometimes more advantageous to use a complex feature together with its simple component features, compared with using simple features alone. For example, if we have two simple features in a model: (A) if a period is preceded by a lowercased word, the period is probably a sentence boundary and (B) if a period is followed by an uppercased word, the period is probably a sentence boundary, the model will perform better, with the addition of a complex feature C combining A and B: (C) if a period is preceded by a lowercased word and followed by an uppercased word, the period is probably a sentence boundary.

Another potential problem when using MEM for SBD is that some useful features may be rarely seen in a particular training set and thus may not be estimated reliably. In such cases, we have tried to group similar situations into one single feature, which has more instances in the training set. For example, instead of using a separate feature to model that a period is unlikely to be a sentence boundary if the next token is "\"", we have modeled "\"" together with several other punctuations as in the following feature:

\[ f(x, y) = \begin{cases} 1 & \text{if } x \text{ denotes a capitalized word following a period,} \\ & \text{and } y \text{ indicates the period is a sentence boundary.} \\ 0 & \text{otherwise} \end{cases} \]

The best model \( p^* \), among all the models derived from the above, is the one maximizing conditional entropy \( H(p) \) on \( p(y|x) \), because it corresponds to a model with more uniform probability distribution on \( x \) unseen in the training set, thus, allowing less bias for unseen contexts.

\[
H(p) = - \sum_{x,y} p(x)p(y|x) \log p(y|x)
\]

\[
p^* = \arg\max_p H(p)
\]
analyzed as either a stream of [adjective-adjective-plural noun-plural noun] or [adjective-adjective-plural noun]-[present -s form of verb]] depending on whether "images" is tagged as a plural noun or as a verb. A lexicalized parser compares alternative parses using the probabilities based first on word–word dependencies, relying on word-free contexts only as a secondary strategy. However, although the word "images" is a common newswire word, the word pair "hemisphere images" does not occur in the WSJ training set. In comparison, an unlexicalized parser does not rely on word–word dependencies, but instead uses the probabilities associated with syntactic categories only. In this model, the word pair "hemisphere images" would be no more anomalous than "brand images," since both involve the same syntactic categories. Unlexicalized probabilities, in general, are less sensitive to the training domain, which may potentially lead to better generalization performance. That is not to say that there are no syntactic differences between newswire text and medical text, just that these differences are less pronounced than lexical differences. The second advantage of this parser is that it reads its grammar and lexicon in a text format that is easily extensible. We exploited that extensibility in this study to adapt the parser to the medical domain by loading in a slightly changed grammar (see Results and Discussion sections for performance improvements and changes we made). We also implemented a NP tagger, which traverses the parse tree and marks up NPs. The Stanford Parser was trained on a nonmedical document collection (The Penn Treebank, WSJ section); thus, many biomedical terms found in clinical documents were rarely encountered by the parser in training. To improve the performance of the parser, we added the following preprocessing (other than sentence delimiting). First, some tokenization improvements were implemented based on sentence delimitation results. Second, the text in some sections of the radiology reports, such as the Impression section, were all in uppercase, which had a very negative impact on the parser’s performance since the parser extensively used lexical features based on different letter cases. Thus, a preprocessing module converted all uppercase texts to proper-cased texts. In this process, our program also detected abbreviations that were not converted to lowercase, using a list of more than 4,000 abbreviations derived from the UMLS Specialist Lexicon. Third, we attempted to customize the frequencies of a few commonly used words, because the statistics learned from the parser’s training set were so different from those in the clinical document set. Last, we constructed an extended lexicon of biomedical terms by mapping the POS tags in the lexical entries of the UMLS Specialist Lexicon to the relevant standard Penn Treebank POS tags. This extended lexicon was then used with the parser to improve the performance of NPI in clinical documents.

The UMLS Specialist Lexicon uses its own syntactic categories to categorize words, which can not be used by the Stanford parser directly. To convert the UMLS Specialist Lexicon entries into new entries with Penn Treebank tags, we used the mapping shown in Table 2. Some of those categories were not mapped because they are closed categories, meaning they have only a certain number of words already well represented by the parser lexicon. Of note, we limited the conversion to unambiguous entries. In other words, within the extended lexicon, we only included those words with only one allowable syntactical category (POS tag), the Stanford statistical parser requires the relative frequency of each tag, which is absent from the UMLS Specialist Lexicon. Second, some mapping conversions were inherently ambiguous. Third, the parser does have a robust mechanism for handling unknown words using frequency information on their lexical features, despite being trained on newswire text. Given the choice between this mechanism and a partial domain lexicon lacking frequencies, it was better to leave the parser unmodified. After conversion, there were 262,704 entries in the extended lexicon, drawn from the UMLS Specialist Lexicon (2004 version), while the original Penn Treebank lexicon contained fewer than 100,000 entries. Thus, the use of the UMLS Specialist Lexicon more than doubled the size of original lexicon.

**Table 1: Sample Features Used in Maximum Entropy Modeling (MEM) Sentence Boundary Detection**

<table>
<thead>
<tr>
<th>Features Used by Maximum Entropy Modeling Sentence Boundary (SB) Detection</th>
<th>Examples: Left Candidate Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left is a lowercased word, Candidate is a SB.</td>
<td>cranietomy. Extracranially</td>
</tr>
<tr>
<td>Right is a lowercased word, Candidate is not a SB.</td>
<td>e. the</td>
</tr>
<tr>
<td>Right is ',', '?', ',', or ' ', Candidate is not a SB.</td>
<td>curvilinear .</td>
</tr>
<tr>
<td>Left is an honorific, Candidate is 'Mrs', Candidate is not a SB.</td>
<td>Dr. Smith</td>
</tr>
<tr>
<td>Candidate is an odd numbered quote, Candidate is not a SB.</td>
<td>&quot; A</td>
</tr>
<tr>
<td>Left is a single uppercase letter and Candidate is 'Mrs', Candidate is not a SB.</td>
<td>S. levels</td>
</tr>
</tbody>
</table>

Sequence of tokens (token1 token2 token3) is represented as (Left Candidate Right).
the UMLS are expressed in simple NPs, for example, as “cerebellar hemisphere” (C0228465). At other times, the most specific UMLS concepts are expressed in longer, more complex, NPs, e.g., “insertion of graft of great vessels with cardiopulmonary bypass” (C0189681). In a full parse tree of a sentence, maximal NPs are more complex in structure and sit closer to the root of the parse tree. Simpler NPs and base NPs are usually nested in the maximal NPs and sit closer to the leaves of the parse tree.

**Outcome Measures**

Precision, recall, and F1 measure\(^49\) were used in evaluating results in this study. Precision is the fraction of proposed NPs, which are present in the gold standard. Recall is the fraction of gold standard NPs that are proposed by our system. We compared performance for statistical significance by calculating 95% confidence intervals for recall and precision using the method provided by Wilson.\(^50\) F1 is a combined measure, defined to be the harmonic mean of these two quantities, which is computed as \(2PR/(R+P)\), where \(P\) is precision and \(R\) is recall. The F1 measure is useful in giving us a single numeric measure of overall performance combining both precision and recall.

**Analysis**

We set out to test three hypotheses:

1. Heuristics could be developed to reliably identify sentences in clinical radiology reports.
2. Using a statistical NLP parser we can accurately identify most NPs in clinical radiology reports.
3. Integrating the UMLS Specialist Lexicon (SL) into an NLP parser trained on nonmedical documents can lead to substantial improvements to NPI in clinical radiology reports.

**Hypothesis 1— Sentence Boundary Detection**

To generate the SBD gold standard, we first wrote a simple rule-based application to pre-markup the sentences in all 100 radiology reports. One of the authors then went through the reports and generated the gold standard of sentence markups by correcting and adding sentence markups. The MEM sentence boundary detector was trained on a training
set of 50 of these reports and then tested on the other 50 reports using this gold standard.

Hypothesis 2—Noun Phrase Identification
Taking parse trees output by the Stanford parser trained on Penn Treebank Wall Street Journal (WSJ) newswire corpus, a NP tagger was developed to identify NPs. The authors performed experiments using both base and maximal NPs, since there is no universal level of complexity at which NPs are optimal for UMLS indexing. One experiment identified and used all maximal NPs in the documents, whereas the other identified and used all base NPs. By looking at both top level and bottom level NPs in parse trees, we hoped to derive a more reliable evaluation of NLP parsing performance, which we believe to be critical to UMLS indexing.

These two experiments were similar except for the preparation of the gold standard and small differences in programs used to tag NPs. To identify maximal NPs correctly, domain knowledge is needed to resolve prepositional phrase attachment and other structure ambiguities; thus, four physicians helped us in creating the gold standard (the process is explained below). On the other hand, the identification of the base NPs is usually much more straightforward for humans; thus, one of the authors was able to create the gold standard in marking up base NPs.

The NPI system processed the 100 reports by first delimiting and parsing all sentences, and then marking up the NPs within each sentence. In the first experiment of identifying maximal NPs, the marked-up reports generated by this process were split into four sets of 25 reports and reviewed by four physicians to identify false-positive and false-negative NP markups. One author then went through all 100 reports and decided the final markups through discussions with the physicians. Based on this expert review, corrections were then made to the markups, and the resulting 100 documents were used as the gold standard of the final NPI results.

In the second experiment of identifying simpler base NPs, one author reviewed the marked-up reports, made corrections to the markups, and the resulting 100 documents were used as the gold standard. We used the same set of 50 reports as in the SBD experiment as the test set in NPI evaluation.

Hypothesis 3—Integrating the UMLS Specialist Lexicon
The computed NP markups with and without the extended lexicon were then compared against the gold standard mentioned in the previous section. Precision, recall, and F1 measure were calculated for the two versions of the NPI module. Because there were published results on base NPI in the general domain,\textsuperscript{31,32} it was possible to compare the base NPI performance of our system with systems working in the nonbiomedical domain. Thus, in addition to the comparison of NPI performance with and without the extended lexicon, we also derived a customized grammar by making a few changes to the grammar learned while training the parser on the Penn Treebank WSJ corpus. Statistical parsers usually use a lexicon with relative frequencies of different syntactical categories for each token. These frequencies are used by parsers to generate the most likely parse for each sentence. However, these frequencies are not present in the Specialist Lexicon. Some commonly used words may have very different relative frequencies in a specific domain other than the general domains. For example, “left” can be the past tense or past/passive participle of the verb “leave,” or simply an adjective that indicates laterality. In the Penn Treebank corpora, “left” is mostly used as a verb. However, in clinical documents such as radiology reports, “left” is almost always used as an adjective to indicate laterality. In base NPI, we manually added new frequencies for a few words such as “left,” “patent,” etc. in the parser grammar. The changes were derived by reviewing reports in the training set only.

Results

Hypothesis 1—Sentence Boundary Detection
The results of sentence boundary detection on the test set of 50 reports are shown in Table 3. The precision was generally excellent (>99%) except for radiology procedure reports (97%). The recall was also very good (>98%) except for radiographs (96%). Overall, our SBD module achieved 99.3% precision with a 95% confidence interval (CI) 98.4-99.7%, 98.3% (97.2-99.0%) in recall and 98.8% in F1 measure.

Hypothesis 2—Noun Phrase Identification
The NP identification module used an unlexicalized version of the Stanford parser to parse reports and identified NPs from parse trees generated by the parser. The developers of the NPI module used the training set to evaluate the performance and improve the module. The NPI module was then run against the test set. The results of the identification of maximal and base NPs using the Stanford parser without any help from a biomedical lexicon are shown in Table 4.

Hypothesis 3—Integrating the UMLS Specialist Lexicon
In the third part of the experiment, we measured the performance of maximal and base NPI with the extended lexicon generated from the UMLS Specialist Lexicon. The results are shown in Tables 5-8.

The results of the identification of maximal NP are shown in Table 5. Without help from the UMLS Specialist Lexicon (baseline), the NPI module achieved 78.9% (76.8% to 80.9%) precision, 81.5% (79.5% to 83.4%) recall, and 80.2% F1

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Results of Sentence Boundary Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report Type</td>
<td>Human</td>
</tr>
<tr>
<td>CT</td>
<td>267</td>
</tr>
<tr>
<td>MAMMO</td>
<td>13</td>
</tr>
<tr>
<td>MR</td>
<td>260</td>
</tr>
<tr>
<td>PROC</td>
<td>131</td>
</tr>
<tr>
<td>RAD</td>
<td>136</td>
</tr>
<tr>
<td>US</td>
<td>19</td>
</tr>
<tr>
<td>All</td>
<td>826</td>
</tr>
</tbody>
</table>
measure overall. Using the extended lexicon (SL), it achieved 82.1% (80.1% to 83.9%) precision, 84.6% (82.6% to 86.3%) recall, and 83.3% F1 measure. The contribution of including terms from the UMLS Specialist Lexicon can be seen more readily in Table 6. The improvements in precision and recall were statistically significant at the 95% confidence level, because the latter precision and recall central values were above the upper bound of 95% CI of the former precision and recall, respectively. Overall, the false-positives and false-negatives were cut by 15.3% and 16.5% by using the extended lexicon.

Similarly, the results of the identification of base NPs are shown in Table 7. Besides the comparison of NPI performance with (SL in the table) and without (baseline in the table) the extended lexicon, we tried to use a customized grammar (SL+GR) to further improve the performance of base NPI. Compared with baseline performance of 86.7% (85.4% to 88.1%) precision, 86.7% (85.4% to 88.1%) recall, and 86.7% F1, using the extended lexicon constructed from the UMLS Specialist Lexicon improved results to 90.9% (89.6% to 92.0%) precision, 91.3% (90.1% to 92.4%) recall, and 91.1% F1. The improvements were statistically significant.

Table 6 shows performance changes, by using an extended lexicon (SL), and using both an extended lexicon and a customized grammar (SL+GR) to further improve the performance of base NPI.
customized grammar (SL+GR), both compared with the baseline. Overall, using the extended lexicon improved the F1 measure by 5.0% and cut the false-positives and false-negatives by 31.1% and 34.4%, respectively. The final NPI module, with both the extended lexicon and a few changes in the grammar, improved the F1 measure by 7.1% and reduced the false-positives and false-negatives by 48.3% and 44.2%, respectively.

**Discussion**

The ultimate goal of the ChartIndex project is to create a CDA-compliant model of clinical document representation at both the structural and semantic levels. The semantic model requires an indexing engine that can automatically identify and then represent important biomedical concepts in clinical documents as UMLS concept descriptors. Prior work in this area has found that achieving good indexing precision is a major challenge. Our current work uses a variety of approaches to address this issue. In this report we show that using a combination of machine learning and NLP can aid in the automated identification of sentence boundaries and NPs in clinical radiology reports.

Most existing NLP systems identify phrases using shallow parsing or text-chunking methods. This is partially because chunking systems are faster than full parsers and partially because full parsers are perceived as error prone.51,52 However, with recent advances in statistical parsing methods, we believe that full parses are now better able to resolve important ambiguities within a reasonable time. It takes 1 to 2 seconds for our parser to parse a sentence of average length of 25 words on a Pentium 4 2.8 GHz computer with 1 GB SDRAM, which is sufficiently fast for our current

---

**Table 7** Results of Base Noun Phrase Identification

<table>
<thead>
<tr>
<th>Report Type</th>
<th>NPI Version</th>
<th>True-Positive</th>
<th>False-Positive</th>
<th>False-Negative</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>Baseline</td>
<td>612</td>
<td>111</td>
<td>98</td>
<td>0.846</td>
<td>0.862</td>
<td>0.854</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td>649</td>
<td>78</td>
<td>61</td>
<td>0.893</td>
<td>0.914</td>
<td>0.903</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>663</td>
<td>49</td>
<td>47</td>
<td>0.931</td>
<td>0.934</td>
<td>0.932</td>
</tr>
<tr>
<td>MAMMO</td>
<td>Baseline</td>
<td>26</td>
<td>7</td>
<td>7</td>
<td>0.788</td>
<td>0.788</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td>29</td>
<td>2</td>
<td>4</td>
<td>0.935</td>
<td>0.879</td>
<td>0.906</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>29</td>
<td>2</td>
<td>4</td>
<td>0.935</td>
<td>0.879</td>
<td>0.906</td>
</tr>
<tr>
<td>MR</td>
<td>Baseline</td>
<td>745</td>
<td>100</td>
<td>105</td>
<td>0.882</td>
<td>0.876</td>
<td>0.879</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td>774</td>
<td>70</td>
<td>76</td>
<td>0.917</td>
<td>0.911</td>
<td>0.914</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>783</td>
<td>57</td>
<td>67</td>
<td>0.932</td>
<td>0.921</td>
<td>0.927</td>
</tr>
<tr>
<td>PROC</td>
<td>Baseline</td>
<td>359</td>
<td>39</td>
<td>31</td>
<td>0.902</td>
<td>0.921</td>
<td>0.911</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td>368</td>
<td>26</td>
<td>22</td>
<td>0.934</td>
<td>0.944</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>370</td>
<td>21</td>
<td>20</td>
<td>0.946</td>
<td>0.949</td>
<td>0.948</td>
</tr>
<tr>
<td>RAD</td>
<td>Baseline</td>
<td>275</td>
<td>47</td>
<td>60</td>
<td>0.854</td>
<td>0.821</td>
<td>0.837</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td>299</td>
<td>35</td>
<td>36</td>
<td>0.895</td>
<td>0.893</td>
<td>0.894</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>301</td>
<td>32</td>
<td>34</td>
<td>0.904</td>
<td>0.899</td>
<td>0.901</td>
</tr>
<tr>
<td>US</td>
<td>Baseline</td>
<td>32</td>
<td>11</td>
<td>12</td>
<td>0.744</td>
<td>0.727</td>
<td>0.736</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td>38</td>
<td>6</td>
<td>6</td>
<td>0.864</td>
<td>0.864</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>42</td>
<td>2</td>
<td>2</td>
<td>0.955</td>
<td>0.955</td>
<td>0.955</td>
</tr>
<tr>
<td>All</td>
<td>Baseline</td>
<td>2,049</td>
<td>315</td>
<td>313</td>
<td>0.867</td>
<td>0.867</td>
<td>0.867</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td>2,157</td>
<td>217</td>
<td>205</td>
<td>0.909</td>
<td>0.913</td>
<td>0.911</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>2,188</td>
<td>163</td>
<td>174</td>
<td>0.931</td>
<td>0.926</td>
<td>0.928</td>
</tr>
</tbody>
</table>

**Table 8** Performance Changes from Baseline by Using Extended Lexicon, and Using Both Extended Lexicon and Customized Grammar

<table>
<thead>
<tr>
<th>Report Type</th>
<th>NPI Version</th>
<th>Change in True-Positives</th>
<th>Change in False-Positives</th>
<th>Change in False-Negatives</th>
<th>Change in Precision</th>
<th>Change in Recall</th>
<th>Change in F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>SL</td>
<td>6.0%</td>
<td>-29.7%</td>
<td>-37.8%</td>
<td>5.5%</td>
<td>6.0%</td>
<td>5.8%</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>8.3%</td>
<td>-55.9%</td>
<td>-52.0%</td>
<td>10.0%</td>
<td>8.3%</td>
<td>9.2%</td>
</tr>
<tr>
<td>MAMMO</td>
<td>SL</td>
<td>11.5%</td>
<td>-71.4%</td>
<td>-42.9%</td>
<td>18.7%</td>
<td>11.5%</td>
<td>15.0%</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>11.5%</td>
<td>-71.4%</td>
<td>-42.9%</td>
<td>18.7%</td>
<td>11.5%</td>
<td>15.0%</td>
</tr>
<tr>
<td>MR</td>
<td>SL</td>
<td>3.9%</td>
<td>-30.0%</td>
<td>-26.9%</td>
<td>4.0%</td>
<td>3.8%</td>
<td>3.9%</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>5.1%</td>
<td>-43.0%</td>
<td>-35.6%</td>
<td>5.7%</td>
<td>5.0%</td>
<td>5.3%</td>
</tr>
<tr>
<td>PROC</td>
<td>SL</td>
<td>2.5%</td>
<td>-33.3%</td>
<td>-29.0%</td>
<td>3.5%</td>
<td>2.5%</td>
<td>3.0%</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>3.1%</td>
<td>-70.0%</td>
<td>-68.3%</td>
<td>9.4%</td>
<td>8.2%</td>
<td>8.8%</td>
</tr>
<tr>
<td>RAD</td>
<td>SL</td>
<td>8.7%</td>
<td>-25.5%</td>
<td>-40.0%</td>
<td>4.8%</td>
<td>8.7%</td>
<td>6.8%</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>9.5%</td>
<td>-31.9%</td>
<td>-43.3%</td>
<td>5.8%</td>
<td>9.5%</td>
<td>7.7%</td>
</tr>
<tr>
<td>US</td>
<td>SL</td>
<td>18.8%</td>
<td>-45.5%</td>
<td>-50.0%</td>
<td>16.1%</td>
<td>18.8%</td>
<td>17.4%</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>31.3%</td>
<td>-81.8%</td>
<td>-83.3%</td>
<td>28.3%</td>
<td>31.3%</td>
<td>29.8%</td>
</tr>
<tr>
<td>All</td>
<td>SL</td>
<td>5.3%</td>
<td>-31.1%</td>
<td>-34.3%</td>
<td>4.8%</td>
<td>5.2%</td>
<td>5.0%</td>
</tr>
<tr>
<td></td>
<td>SL+GR</td>
<td>6.8%</td>
<td>-48.3%</td>
<td>-44.2%</td>
<td>7.4%</td>
<td>6.7%</td>
<td>7.1%</td>
</tr>
</tbody>
</table>
applications. The performance of the parser was evaluated on its native training domain in a previous study, although the accuracy on medical texts has not been explicitly tested. A full parser also offers the ability to predict larger NPs that most text chunkers do not attempt to predict. Moreover, full parsers make more detailed assertions about relational syntactic structures, which can reasonably be expected to be useful for future indexing work.

Another concern of applying a statistical parser to a domain other than its training native domain is performance degradation. In this study, we applied the Stanford parser, trained on the Penn Treebank WSJ corpus, to clinical radiology documents in the medical domain. As Gildea reported, word-to-word dependencies are corpus specific. Thus, the unlexicalized Stanford parser gave us a compact and fast statistical parser that we believe is less dependent on its training corpus. Furthermore, we augmented the Stanford parser with a biomedical lexicon derived from the UMLS Specialist Lexicon. The performance of the parser on our document collection, although not evaluated on full parse trees, was evaluated at the levels of both maximal and base NPs. The finding by Hwa that higher-level constituents are the most informative linguistic units in grammar induction suggests that the evaluation on maximal NPI might be a better indicator than base phrases of overall parsing accuracy. As shown above, the results of maximal NPI are acceptable, and those of base NPI are comparable to published performance in the newspaper, which is the parser’s native training domain.

In addition, we believe that the representation of NPs within a parse tree provides considerable potential flexibility at the time of indexing. Compared with the flat output structure produced by text chunking, a parse tree captures more structural information revealing the semantics of the sentence, which may be very helpful in identifying negated concepts. This approach can support heuristics that select the optimal NP node for indexing by traversing paths between the maximal NPs and base NPs within the parse tree.

Noun phrase identification is a critical step in the ChartIndex model. Most important biomedical concepts in clinical documents are NPs, and most UMLS concept descriptors are NPs. The identification of NPs in ChartIndex relies on a high-performance statistical parser. High performance statistical parsers are usually trained on corpora in a general domain. To apply those parsers to clinical documents effectively, we need to supply the parser with biomedical terms. We have shown the UMLS Specialist Lexicon to be an effective resource for this purpose. As mentioned above, there are generally two problems in integrating SL with these parsers. The first one is the mismatch of syntactical categories, which causes ambiguities in mapping entries in SL syntactical categories to entries in Penn Treebank syntactical categories. To address this issue we chose to do the mapping conservatively, by mapping only unambiguous terms. A second problem is that statistical parsers usually use a lexicon with relative frequencies of different syntactical categories for each token. Those relative frequencies for some common words may be very different in biomedical domain from those in general domains. In the base NPI experiment, we manually changed relative frequencies for a few words as mentioned in the Methods section. Those results are marked as “SL+GR” in the data tables. Table 7 shows an improvement from 91.1% F1 to 92.8% F1 for all reports. While this tuning is extremely specific, and might raise concerns of overfitting, certain words are so extremely common and so extremely different in distribution between domains that even just a few such modifications (for five words in this case) can be widely applicable and generally useful.

The experiment on the identification of maximal NPs reported consistently lower performance numbers, with F1 of 80.2% without terms from the UMLS Specialist Lexicon and F1 of 83.3% with terms from the UMLS Specialist Lexicon. One reason is as mentioned above, to identify the maximal NP is indeed a more difficult problem because the parser needs to resolve attachment ambiguities. Another reason is related to how the errors are counted. For example, “occluded left FEM to distal bypass graft” is a maximal NP in a sentence. The parser mistakenly marked “left” as a verb, which led to three false-positives of “occluded,” “FEM,” and “distal bypass graft” and one false-negative, in the maximal NPI. On the other hand, in base NPI, the same parse error only led to two false-positives, “occluded” and “FEM,” since “distal bypass graft” is a correct base NP. Additionally, because the total number of maximal NPs is smaller than the total number of base NPs, a failure is weighted more in maximal NPI results.

From Table 5, we can see that the performance of maximal NPI has less variation across different modalities after applying the extended lexicon, except ultrasounds (US), which may reflect the small data set. From Table 6, we can see that adding terms from SL improved performance consistently, except in the smaller data set of mammogram (MAMMO), in which there were no changes in performance. Across the whole test set, there was a reduction of 15.3% in false-positives and 16.5% in false-negatives with the use of the Specialist Lexicon.

Table 7 shows the same trend in base NPI. The baseline F1 measure ranges from 73.6% in ultrasound (US) to 91.1% in radiology procedure (PROC). With SL terms, the F1 measure was improved to between 86.4% in US and 93.9% in PROC. The final version using both SL terms and a grammar with a few changes further improved F1 to between 90.1% in radiograph (RAD) and 95.5% in US. Due to the factors mentioned above, we can see the performance improvements are more substantial in base NPI by comparing the data in Table 8 with the data in Table 6.

There were some challenges presented by this document set. For example, as mentioned previously, some sentences were not well formed. Most of the ill-formed sentences were NPs. The parser could parse most of them correctly but had some problems with some long complex NPs in the Impression section. Sometimes those NPs were parsed as a full sentence either with a NP and a verb phrase or with a verb phrase only. In both cases, it was usually because some words in the text have more than one POS tag, and they were not pretagged using the extended lexicon. These words were also rare words (seen by the parser less than 20 times in the whole training process) and were tagged as a verb and the head of a verb phrase by the parser. The second type of error originated from commonly used words in capitalized form, such as “Right.” The parser currently treats “Right” as a separate entry from “right” since “Right” may
serve as part of a proper noun. There were also a few cases of parsing errors involving punctuations like parenthesis and the pound sign "#". These errors were due to the fact that these punctuations are used differently in radiology reports and indicate that there are syntactic adaptations that still remain to be done. These cases could not necessarily be addressed by simple lexicon adaptation.

Statistical learning techniques have been widely adopted in text processing and text mining applications and have been shown to produce robustness and good performance. One potential hurdle associated with this approach is the need for large-labeled training sets for use in supervised machine learning. This is especially true in the biomedical domain since there are few publicly available large-labeled corpora of clinical documents, such as the Penn Treebank corpora for general domains. We have extended such a parser using a domain-specific lexicon. There are still some remaining issues with this approach such as the ambiguous tag mappings in the conversion of lexicons using different POS tag sets. Also, the probabilities associated with POS tags for each term can be very different in clinical documents compared with the training collection in a general domain. However, our work has shown that, with the help of lexical entries contained in a domain-specific lexicon (the UMLS Specialist Lexicon) a statistical natural language parser trained on a training set in the general domain can achieve significantly improved performance on NP identification within clinical radiology reports.

There are limitations to our analysis. First, the method that we used to create a gold standard was not optimal. Ideally we would have asked each expert physician to review all 100 radiology reports. However, this was not possible given the time constraints of these experts and we instead asked each physician to review 25 of the 100 documents. If all 100 documents had been reviewed by each physician, we could have evaluated per-rater reliability and intra-rater variability. Second, the use of computer pre-markups may bias human experts’ judgments. Third, we did not evaluate the entire parse tree for each sentence, although we did get a good estimate of the parser’s overall performance through evaluation of the base NPI and maximal NPI.

Conclusion

The performance of sentence boundary detection is excellent in this system. Extraction of NPs in clinical radiology reports, using statistical natural language processing, can achieve performance comparable to that seen in the general, nonmedical, domain. The adaptation using the UMLS Specialist Lexicon significantly improved both precision and recall in NPI on clinical radiology reports to levels comparable to the parser’s native performance in its nonbiomedical training domain (newswire). Future work will include the development of a system that will take NPs in parse tree format and map them into corresponding UMLS concepts.

References


