Proceedings of the 2007 IEEE Symposium on
Computational Intelligence and Data Mining (CIDM 2007)

Identifying Anatomical Phrases in Clinical Reports by Shallow Semantic Parsing Methods

Vijayaraghavan Bashyam, Ricky K Taira
Medical Imaging Informatics Group,
University of California, Los Angeles
924 Westwood Blvd, Suite 420, Los Angeles, CA 90024
vijay | rtaira @ mii·ucla·edu

***Abstract*— Natural Language Processing (NLP) is being applied for several information extraction tasks in the biomedical domain. The unique nature of clinical information requires the need for developing an NLP system designed specifically for the clinical domain. We describe a method to identify semantically coherent phrases within clinical reports. This is an important step towards full syntactic parsing within a clinical NLP system. We use this *semantic phrase chunker* to identify *anatomical phrases* within radiology reports related to the genitourinary domain. A discriminative classifier based on support vector machines was used to classify words into one of five phrase classification categories. Training of the classifier was performed using 1000 hand-tagged sentences from a corpus of genitourinary radiology reports. Features used by the classifier include n-grams, syntactic tags and semantic labels. Evaluation was conducted on a blind test set of 250 sentences from the same domain. The system achieved overall performance scores of 0.87 (precision), 0.91 (recall) and 0.89 (balanced f-score). Anatomical phrase extraction can be rapidly and accurately accomplished.**

Keywords- Natural language processing, shallow semantic parsing, anatomy phrases, radiology reports, support vector machines

I. INTRODUCTION

The adoption of the electronic medical record by hospitals in the United States has resulted in the generation of large volumes of textual data on an everyday basis as a result of routine clinical care. This information is largely in the form of unstructured natural language [1, 2]. The structuring of such narrative reports is vital for using the rich information contained in them such as descriptions of the state of disease. The need for developing tools to extract such information from biomedical text has often been stressed in the past [3]. We have developed a method to identify semantically coherent phrases within medical reports as an important step towards full syntactic parsing. Using this technique, we attempted to mine *anatomical phrases* from radiology reports within the genitourinary domain. Anatomical phrase identification is of utmost importance for clinical natural language processing (NLP) because clinical reports primarily consist of anatomical concepts associated with other concepts. For example, a radiology report contains descriptions of findings in anatomical locations. A surgery report contains a description of actions performed on anatomical parts. Correctly identifying anatomy phrases is also an important step towards coding concepts to a standard vocabulary.

This paper reports the performance of this NLP system to identify anatomical phrases in urology related radiology reports. The remainder of this paper is organized as follows. Section II provides a brief background on the need for semantic phrasal chunking and reviews the previous methods used for similar tasks. Section III describes the problem formalization, data collection and implementation of the methods. Section IV summarizes the results of this experiment and Section V concludes with an error analysis and future directions for this project.

II. BACKGROUND

A. Need for Semantic Phrase Chunking

'Phrase chunking' can be defined as the identification of logically coherent non-overlapping sequences of words within a sentence. It is an intermediate step towards full syntactic parsing [4] and is an effective method of reducing the dimensionality of the overall NLP task. Traditionally phrase chunking has been primarily syntactic in nature. In other words, the boundaries of the phrase being extracted are marked according to grammars so that the resulting phrase conforms to an established syntactic structure. The most frequently occurring phrases in this scheme of phrase chunking are noun phrases, verb phrases and prepositional phrases [5]. The phrase boundaries are usually marked according to the conventions followed by the Penn Tree Bank [6].

The disadvantage of using syntactic chunking (at least) in the clinical domain is due to the difficulty in obtaining grammatically correct sentences in medical reports. This problem has been acknowledged in the domain of clinical pediatric literature [7] and to a lesser extent, even in the domain of general language [8, 9]. Physicians often dictate their diagnosis to a speech recognition system which transcribes their dictation to text. Though the physician manually inspects the report to correct it for transcription errors, it is uncommon to find medical reports with strict punctuation. Peculiarities like homophones and abbreviations are sources of noise in automated / manual transcription. Physicians often use partial sentences (e.g. 5cm mass seen,)

This work was supported in part by the following grants: 1. National Institute of Biomedical Imaging and Bioengineering NIBIB P01-EB00216, 2. National Institute of Health RO1-EB002247

and ungrammatical constructions. Such issues make conventional NLP systems (trained in the domain of newspaper text), incompatible with medical text unless suitable modifications are made such as integrating a medical lexicon to a tree-bank parser [10].

In addition, not all semantically coherent phrases found in medical reports can be identified with syntactic phrases. Clinical information is inherently complicated in nature and it is important to cluster the right set of words together so that the semantics of the information is preserved. Sometimes a phrase which may be semantically coherent could be split across syntactic phrases. "Fig. 1" shows the syntactic parse of a sentence containing a spatial relation phrase *'is located just inferior to'* split across multiple syntactic phrases.

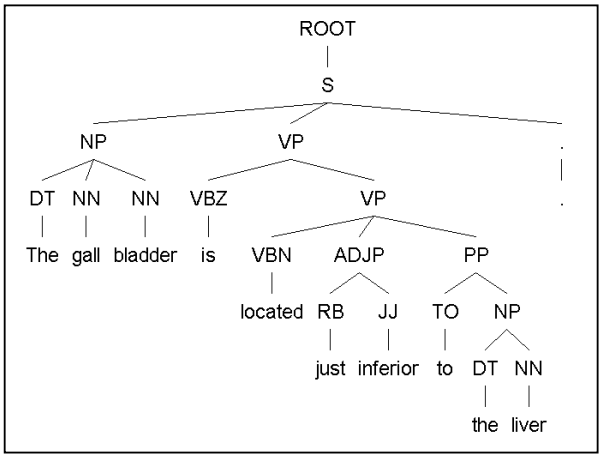

Figure 1. Parse Tree obtained from the Stanford Parser v1.5.1[11, 12]

Noun phrases are the most common kind of syntactic phrases associated with medical terminology [13]. Given a parse tree structure, it can be difficult to estimate which noun phrase in the parse hierarchy is ideal to be categorized as a single coherent unit [10].

Such disadvantages require the need for clustering words differently in clinical text. We define a *semantic phrase* to be a sequential set of word tokens which can be effectively replaced by a single word belonging to the same semantic category as the phrase. Thus, an *anatomical phrase* is a continuous set of words which can be effectively replaced by a single anatomical word. For example consider the sentence: "There are masses seen in the left occipital lobe, portions of the temporal lobe and frontal lobe." The anatomical phrase here is *left occipital lobe, portions of the temporal lobe and frontal lobe*.

B. Previous Methods

The early methods in chunking focused primarily on identification of syntactic phrase chunks in general and noun phrases in particular. Such methods include grammar based methods, statistical methods based on the frequency of occurrence of words or their part-of-speech (POS) tags, and classifier based methods.

Bourigault [14] developed a heuristics based system to extract maximal length noun phrases from French text . He then used a grammar to split these phrases into smaller chunks. Bourigault reported that his system could break 95% of the maximal length noun phrases into smaller noun phrases. However his evaluation was not very systematic. There is no report of false positives or conventional precision/recall statistics. Voulilainen [15] described a rule-based chunking system called NPtool which used morphological and syntactic annotations. The system achieved a recall of over 98.5% and precision of 95-98%. Church [16] trained a statistical model on a corpus to identify noun phrase chunks by inserting brackets into text. This system called the Parts program reported qualitative results as 'very encouraging'. The most well known study in chunking is the one carried out by Ramshaw and Marcus [17]. The system based on a large set of transformation rules, achieved over 90% precision and recall.

Until 2000, attempts were focused primarily on noun phrases. The Computational Natural Language Learning conference CoNLL-2000 had a shared task on chunking where a standard dataset was created so that different methods could be trained on the same data and tested on the same data. Thus the results of different approaches could be easily compared.

Veenstra and Van den Bosch [18] achieved a precision of 91% and a recall of 92% by using memory based learning methods for chunking. They also reported that they achieved the best results using only POS tags as features. Osbourne [19] used a maximum entropy approach for POS tagging as described by Ratnaparkhi [20], and obtained an overall accuracy of 94.88% using a combination of words and POS tags. Johansson [21] used only POS tags in his system based on a maximum-likelihood approach, to achieve a precision of 86% and a recall of 88%

Kudo and Matsumoto's system [22] used support vector machines to differentiate between words that are part of a phrase chunk and those outside of a phrase chunk. Their system which used words, POS tags and contextual words with their tags achieved results of 93% precision and recall. Koeling [23] achieved close results of 92% precision and recall using similar features input to a maximum entropy classifier.

Zhang, Damerau and Johnson [24] described a text chunking system using a generalization of the winnow algorithm. They use a rich set of 520,000 features with around 88 non-zero features for each data point. Their system achieved a performance of 94% precision and recall.

The only work related to anatomical term mining to our knowledge is the study conducted at the National Library of Medicine (NLM). Sneiderman, Rindflesch and Bean [25] use the NLP tools developed at the NLM to identify *'coronary artery'*-associated terminology found in coronary catheterization reports. They report a recall of 83% and a

precision of 88% for the task of identifying the specified anatomical terminology.

To summarize, most previous methods for syntactic chunking used a small feature set limited to words, their POS tags, contextual words and their POS tags and were able to achieve good performance.

III. METHODS

A. Problem Formalization

We model the problem of chunking as a classification problem where each word needs to be tagged with a label which indicates whether or not it is a part of the anatomical phrase. We utilize the 5-class tagging scheme described by Kudo and Matsumoto [18]. The goal of our classifier is to tag each word token in the sentence with one of the following five outcomes: a) Begin [B], b) End [E], c) Inside [I], d) Single [S], and e) Outside [O]. The working definition for each outcome is given in Table I. For example, in the sentence, *"A chest mass in the right upper lobe is seen,"* the markup for the anatomy description phrase is as seen in "Fig. 2".

TABLE I. DEFINITION OF CLASSIFIER OUTCOMES

Class	Definition
B	token is Beginning of a phrase consisting of more than one token
E	token is the End of a phrase consisting of more than one token
I	token is between the start and end of a phrase consisting of more than two tokens
S	token is the lone token of a phrase consisting of only one token
O	Current token is outside of the phrase

A	chest	mass	in	the	right	upper	lobe	is	seen
O	S	O	O	O	B	I	E	O	O

Figure 2. Class markup for individual tokens in a sentence

We used support vector machines (SVMs) as the classifier for our task. SVMs are primarily binary classifiers but can also be used for multi-class problems. SVMs have been previously demonstrated to be extremely accurate for the tasks of syntactic chunking [22], dependency parsing [26] and text categorization [27, 28].

B. Data Collection

In any pattern classification task, it is desirable to compile a large number of quality training examples which reflect the underlying distribution of the pool statistics. A representative sample of training data is important since the training examples reflect exactly how the classifier will behave. Any inconsistencies or errors in tagging can cause significant performance degradation. Thus, decisions have to be made on how to handle somewhat ambiguous tagging assignments such as: *"Left native kidney," "Right true pelvis," "Loops of presumed colon," "On the right side, the femur."* Additionally, there are many instances within medical text of partial descriptions (ellipsis) that require some prior knowledge either expressed within a previous portion of the text or simply understood within the domain. For example the word "tip", "end", and "apex" may or may not refer to some landmark on an anatomical organ.

The following steps were followed to identify sentences with anatomical phrases:

1. Over twelve thousand radiology reports related to urology within the existing hospital database at our institution were captured using an XML-based gateway [29]. These reports were associated with all radiological modalities including magnetic resonance imaging, computed tomography, ultrasound, fluoroscopy, and plain film radiography.
2. Section boundary detection was performed on the reports to break up a report into individual sections such as HEADER, HISTORY, FINDINGS, CONCLUSIONS etc. Following this, sentence boundary detection was performed on the sections. Both of these modules have been previously tested, with recall and precision accuracies of over 99% within the domain of radiology [30].
3. A lexical analyzer processed each sentence performing tokenization, part-of-speech tagging and semantic class tagging. Our lexicon categorizes tokens into twenty syntactic categories and over three hundred semantic categories providing improved discrimination for tasks such as syntactic parsing and semantic interpretation.
4. Word-sense disambiguation was performed on commonly occurring words with very different meanings (*e.g.,* 'T1' as an anatomy entity or a signal type in magnetic resonance imaging). Also, recognition of dates, measurements, and special symbols (e.g., tumor staging classifications) was performed in this step. Deidentification of the reports was performed at this stage [31].
5. Using a high-recall sentence-level phrase parser, all possible anatomical phrase instances within a sentence were conservatively identified rejecting sentences that obviously have no candidates. For example, the filter would reject sentences that do not have at least one word from the following semantic classes: *selfReferenceLocation* (e.g., neck), *physobj.anatomy* (e.g., lung).
6. A domain expert with familiarity to NLP, examined the sentences and hand-tagged the anatomy phrases. The tagged set was verified by a second expert and then stored into a training database to serve as the gold standard for testing and development. On ambiguous terms the domain experts came to a consensus.

With this approach, we collected 1250 sentences and tagged them for anatomy phrases. We then set aside 250 randomly selected sentences for testing and used the remaining 1000

sentences for training the classifier. The 80-20 ratio is in accordance to the standard followed in the CoNLL-2000 shared task on chunking[13].

C. Implementation

We used the SVM^{light} implementation of SVMs freely available from the website http://svmlight.joachims.org/. The input to the classifier is the word to be labeled followed by a set of features. In this case, the features included syntactic tags, semantic labels, and contextual words with their syntactic and semantic labels.

The classifier categorized each word into one of the five target categories. The output of the classifier was compared to the gold standard previously created by the domain experts.

IV. RESULTS

Table II summarizes the performance of the individual class assignments output by the phrase chunker on the 250 sentences from a corpus of genitourinary radiology reports from our institution. The performance on this task is quantified with three rates: 1) precision - the percentage of detected phrases that are correct; 2) recall - the percentage of phrases in the data that were found by the chunker and; 3) balanced f-measure – the weighted harmonic mean of precision and recall. These measures are related to true positive (TP), false negative (FN), and false positive (FP) statistics as follows:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

TABLE II. INDIVIDUAL CLASS PERFORMANCE MEASURES

Class	No.	TP	FP	FN	Recall	Precision	f-score
B	307	263	45	51	0.83	0.85	0.84
I	601	592	64	33	0.94	0.90	0.92
E	307	293	48	61	0.82	0.85	0.84
O	1596	1467	94	107	0.93	0.93	0.93
S	116	87	11	13	0.87	0.88	0.87
Total	2927	2702	262	265	0.91	0.91	0.91

TABLE III. PHRASE IDENTIFICATION PERFORMANCE MEASURES

No. of Phrases	TP	FP	FN	Recall	Precision	f-score
423	350	51	31	0.91	0.87	0.89

Of the 423 anatomical phrases present in the test set, 350 phrases were identified correctly. Of these, 263 phrases were multiword phrases and 87 phrases were single word phrases. The overall precision for identifying anatomical phrases was 87% and the recall was 91% as shown in table III. The overall precision and recall for assigning class labels to the tokens were both 91%. The individual class assignment performances are shown in table II.

V. DISCUSSION

We present a system that is based on discriminative pattern recognition methods to accurately locate anatomical phrases found in clinical text reports. Examples of false positive errors included phrases that were abnormally truncated (e.g., ellipsis) such as the example below:

> "The residual barium which was previously seen in dilated loops of small has appeared to have passed."

Examples of false negatives errors included phrases that were part of a conjunctive phrase such as the example below:

> "Following catheterization of both the urinary bladder and a vaginal orifice, the urinary bladder and vaginal were opacified."

We also note that the system performance for the class assignments [B], [E] and [S] are lower than the other classes. This is expected because it is more difficult to tag the phrase boundaries than to tag the inner words of the phrase. However since a large number of tokens are either inside[I] or outside[O] of a phrase boundary, the overall performance measures still show a high performance. We recognize that the interpretation of the individual class performance scores is more important than the overall performance scores.

Future directions for the project include incorporating more contextual features, and training the classifier to recognize other types of phrases like findings, spatial relations, temporal relations, causal relations, existential relations, etc. Additionally, explorations of the system's adaptability to new clinical domains outside of radiology and urology will be conducted.

VI. CONCLUSION

A fast accurate anatomy phrase parser has been developed within the application area of genitourinary radiology reports. High system accuracy is achieved by a combination of a large number of domain specific training examples, a rich set of discriminating features, and a powerful discrimative classifier. This system will be used both as a part of an NLP system as well as a standalone application to mine anatomical phrases.

PERMISSIONS

This project has been approved by the University of California, Los Angeles, Institutional Review Board (IRB) vide IRB approval number G0012001-13

ACKNOWLEDGMENT

The authors thank Carol Demise and the domain experts for their effort in preparing the anatomy training corpus. The authors especially thank Drs. Hooshang Kangarloo and Paul S. Cho for their guidance in this project.

REFERENCES

[1] H. J. Tange, H. C. Schouten, A. D. M. Kester, and A. Hasman, "The Granularity of Medical Narratives and Its Effect on the Speed and Completeness of Information Retrieval," *J Am Med Inform Assoc,* vol. 5, pp. 571-82, 1998.

[2] F. Hall, "Language of the radiology report," *American Journal of Roentology,* vol. 175, pp. 1239-1241, 2000.

[3] S. Ananiadou, C. Friedman, and J. Tsujii, "Introduction: named entity recognition in biomedicine.," *Journal of Biomedical Informatics,* vol. 37, pp. 393-5, 2004.

[4] S. Abney, "Parsing by chunks," in *Principle-Based Parsing. Kluwer Academic Publishers*, 1991.

[5] E. F. T. K. Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task: chunking," *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7,* pp. 127-132, 2000.

[6] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: the penn treebank," *Computational Linguistics,* vol. 19(2), pp. 313-330, 1993.

[7] J. Pestian, L. Itert, and W. Duch, "Development of a Pediatric Text-Corpus for Part-of-Speech Tagging," in *Intelligent Information Processing and Web Mining*, S. T. W. M.A. Klopotek, K. Trojanowski, Ed., 2004, pp. 219-226.

[8] E. Sapir, *Language: An Introduction to the Study of Speech*: Dover Publications, 2004.

[9] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*: MIT Press, 1999.

[10] Y. Huang, H. J. Lowe, D. Klein, and R. J. Cucina, "Improved Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon," *Journal of the American Medical Informatics Association,* vol. 12, pp. 275-285, 2005.

[11] D. Klein and C. D. Manning, "Fast exact inference with a factored model for natural language parsing," *Advances in Neural Information Processing Systems,* vol. 15, pp. 3-10, 2003.

[12] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* pp. 423-430, 2003.

[13] E. Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task: chunking," in *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*, Lisbon, Portugal, 2000, pp. 127-132.

[14] D. Bourigault, "Surface grammatical analysis for the extraction of terminological noun phrases," *Proceedings of the 14th conference on Computational linguistics-Volume 3,* pp. 977-981, 1992.

[15] A. Voutilainen, "NPtool, a detector of English noun phrases," *Proceedings of the Workshop on Very Large Corpora,* pp. 48-57, 1993.

[16] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text," *Proceedings of the Second Conference on Applied Natural Language Processing,* vol. 136, 1988.

[17] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Proceedings of the Third ACL Workshop on Very Large Corpora*, June 1995, pp. 82-94.

[18] J. Veenstra and A. van den Bosch, "Single-Classifier Memory-Based Phrase Chunking," in *Proceedings of the Fourth Workshop on Computational Natural Language Learning (CoNLL 2000)*, Lisbon, Portugal, 2000, pp. 157–159.

[19] M. Osborne, "Shallow parsing as part-of-speech tagging," *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7,* pp. 145-147, 2000.

[20] A. Ratnaparkhi, "A maximum entropy part-of-speech tagger," in *Proceedings of the Empirical Methods in Natural Language Processing*, University of Pennsylvania, 1996, pp. 133–142.

[21] C. Johansson, "A Context Sensitive Maximum Likelihood Approach to Chunking," in *Proceedings of the 4th Conference on Computational Natural Language Learning -2000*, Lisbon, Portugal, 2000, pp. 136–138.

[22] T. Kudo and Y. Matsumoto, "Chunking with support vector machines," in *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics*, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2001, pp. 192-199.

[23] R. Koeling, "Chunking with Maximum Entropy Models," *Proceedings of CoNLL-2000 and LLL-2000,* pp. 139-141, 2000.

[24] T. Zhang, F. Damerau, and D. Johnson, "Text chunking based on a generalization of winnow," *Journal of Machine Learning Research,* vol. 2, pp. 615-638, 2002.

[25] C. A. Sneiderman, T. C. Rindflesch, and C. A. Bean, "Identification of anatomical terminology in medical text," in *Proceedings of the American Medical Informatics Association Symposium*, 1998, p. 32.

[26] H. Yamada and Y. Matsumoto, "Statistical dependency analysis with support vector machines," in *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, 2003, pp. 195–206.

[27] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of ACM-CIKM98*, 1998, pp. 148-155.

[28] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*: Kluwer Academic Publishers, 2002.

[29] A. A. T. Bui, J. D. N. Dionisio, C. A. Morioka, U. Sinha, R. K. Taira, and H. Kangarloo, "DataServer: An Infrastructure to Support Evidence-based Radiology," *Acad Radiology,* vol. 9, pp. 670-678, 2002.

[30] R. K. Taira and S. G. Soderland, "A statistical natural language processor for medical reports," in *Proceedings of the Amedican Medical Informatics Association Symposium*, 1999, p. 4.

[31] R. K. Taira, A. A. T. Bui, and H. Kangarloo, "Identification of patient name references within medical documents using semantic selectional restrictions," in *Proceedings of the Amedican Medical Informatics Association Symposium, 2002*, 2002, p. 61.