A Statistical Natural Language Processor for Medical Reports

Ricky K. Taira, PhD[1,2] and Stephen G. Soderland, PhD[1,3]

[1]Department of Radiology,
Children's Hospital and Regional Medical Center, Seattle WA 98105

[2]Department of Radiology,
University of Washington, Seattle WA 98105

[3]Department of Radiological Sciences,
University of California, Los Angeles 90024

ABSTRACT

Statistical natural language processors have been the focus of much research during the past decade. The main advantage of such an approach over grammatical rule-based approaches is its scalability to new domains. We present a statistical NLP for the domain of radiology and report on methods of knowledge acquisition, parsing, semantic interpretation, and evaluation. Preliminary performance data are given. A discussion of the perceived benefit, limitations and future work is presented.

INTRODUCTION

To date, most medical natural language processors (NLPs) have been implemented using symbolic methods based on a combination of syntactic and semantic grammars [1-6]. The reported accuracy of these systems has been reasonable within focused domains [7]. However, the effort and adaptability of to new medical domains have yet to be fully investigated [8]. Deployment of NLP systems into clinical environments has been sparse although several preclinical evaluations have been reported [2-8].

Statistical natural language processors (NLPs) [9-15] have recently gained much attention because of some fundamental deficiencies in symbolic methods. Firstly, it is unlikely that one can model language using solely a rule-based system. The rules will never be exhaustive. Most are not absolute, nor independent hence multiple rules can interact poorly. Secondly, the goal of maximizing coverage while minimizing resultant ambiguity is fundamentally inconsistent with symbolic NLP systems [9]. Extending the coverage of the grammar to obscure constructions often leads to an increase in the number of undesired parses. Thirdly, it is time-consuming and difficult to manage the rule base in symbolic systems. In general, rule-based systems do not scale well. Finally, rule based system do not adapt well to unseen patterns or changes in language.

Statistical NLP methods are logical to apply since most tasks in NLP are classification problems [15]. For example, the classification of:

- A period as an end of sentence marker or not,
- A word into its part of speech class
- A link between words as a true dependency or not.

In this paper, we present our initial experience with building a statistical NLP system for radiology reports. We focus on the specific sub-problems of sentence parsing and semantic interpretation.

METHODS

The architecture of our system is similar to many NLP designs and consists of the following modules.

1. ***Structural Analyzer***: Isolates sections of medical reports (e.g., "Procedure Description", "History", "Findings", "Impressions") and individual sentences within sections.
2. ***Lexical Analyzer***: Looks up semantic and syntactic features of words in a medical lexicon [16], normalizes dates and numerical expressions, and tokenizes punctuation.
3. ***Parser***: Determines the dependencies between words. The parser adds arcs that indicate a modifier relationship between pairs of words.
4. ***Semantic Interpreter***: Interprets the links of the parser's dependency diagram and outputs a set of logical relations that form a semantic network for the sentence.
5. ***Discourse Processor***: Determines whether a finding from a sentence is new or a referent to a finding from previous sentences.

1091—8280/99/$5.00

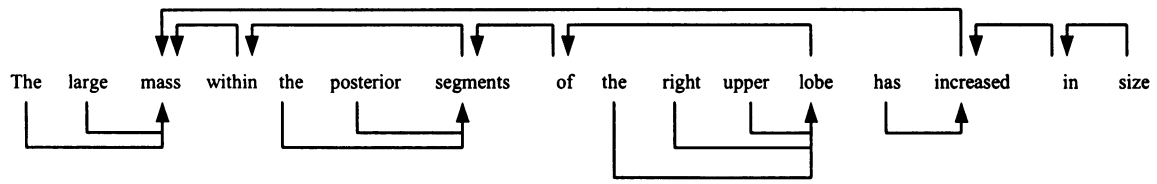

Fig. 1 - *Example sentence with arcs showing dependencies between words.*

The parser and semantic interpreter stages are based on statistical methods and are the focus of this paper. The lexical analyzer currently does not use statistical methods although future plans include implementation of a statistical word-sense disambiguation algorithm [9]. The structural analyzer was recently converted from a rule-based system to one that uses a maximum entropy classifier [15] that uses 40 overlapping features and trained on 6500 sentences.

Parser

The goal of our parser is to create a dependency diagram between words in an input sentence (Fig. 1). An arc from word A to word B indicates A modifies B. We conceptualize the mechanism of parsing as a dynamics problem similar to how atoms aggregate to form complex molecules. This paradigm conceptualizes the process as follows:

- Words initially have no dependencies with other words. They each exist in a *free state*. The free state of a word however, is often not its ideal steady-state. For example, an adjective is unstable without an attachment to an appropriate head noun.
- As the parsing step proceeds, each word attempts to configure itself into a more favorable steady state of existence. We conceptualize forces existing between words indicating word affinities.
- The final state of the parse reflects the configuration of the words that minimizes the overall energy of the system.

To fit our implementation with this paradigm, we designed the system with the following requirements:

- Words must have a mechanism for communicating their identity to other words within the sentence.
- The affinity of a word for other words within the context of the sentence needs to be estimated. For example, which kinds of words can be attached to the word "*large*" in the example in Figure 1.
- A word must be an active entity since in exists within an evolving environment corresponding to different stages of the parse. Words are in constant search for a steady state of existence.

Processor Model for Words

The first two requirements for a word suggest a signal processor model is appropriate. Words then are active entities characterized by their signal processing behavior. This includes its emission spectrum, its absorption spectrum, and its response function to resonance conditions. Indistinguishable words are characterized by their identical signal processing behavior and hence dynamic behavior.

The emission spectrum of a word is represented by features that encode syntactic and semantic information. For example, the word "*mass*" in Figure 1 has syntactic features *noun*, *singular* and has semantic features *abnormal*, *physical object, finding,* and *lesion*. The word "*increased*" has syntactic feature *past participle* and semantic features *temporal, change,* and *increase*.

word	syntax	semantics
the	det	defin_art
mass	noun.sing	abnorm.physobj,finding.lesion
lobe	noun.sing	anat.physobj.struct
increased	pastp	temporal.change.increase

Word-Word Interactions

Given a model for a word, we now need to describe word-word interactions. We again take a signal processor approach describe word-word interactions, conceptualizing this process as a pair of sending and receiving antennas.

First, let's look at the possible interactions that can occur within a simple isolated two-word system. In the discussion that follows, word 'A' has the role of a "signal sender" and word 'B' has the role of a "signal receiver". Word A transmits its characteristic signal (i.e., its emission spectra) towards the receiving target word B. Word B has an inherent absorption spectrum "tuned" to receive only certain types of signals. For example, the word "*mass*" may be tuned to receive signals for words that emit "size" or "shape" information.

The probability of an absorption event occurring between words A and B (i.e., a resonance condition) is determined empirically through supervised learning methods. The methodology is as follows:

- Collect a large sample of documents from the domain of interest (thoracic radiology).
- Create training data by manually indicating the dependency diagram for each sentence to reflect the output from an ideal parser.
- For every pair of words in each sentence, collect statistics on how often a resonance condition occurred vs. not occurred. This is equivalent to an arc existing between words in the training data.

The result of the training set is a table containing statistics for each pair of words (A, B) over all encountered contexts. The statistical distribution may be heavily biased towards resonance or non-resonance.

Take the word "*increased*" in Figure 1 as particle A and consider two possible B particles, "*mass*" or "*lobe*". There is high resonance between "*increased*" and "*mass*", since lesions are frequently described as *increasing (in size)*. The probability of an attachment to "*lobe*", on the other hand, is low. Lobes and other anatomical structures are less frequently described as increasing in thoracic radiology reports.

Estimating resonance statistics from a finite set of training data leads to the problem of sparse data. The number of combinations of all possible words A and B that could co-occur in a sentence is extremely large. A new sentence is likely to have pairs of words that have never been seen in a training set. Two methods of smoothing the statistics are used:

1. Use the semantic and syntactic features of the words instead of the words themselves. The number of unique semantic classes is on the order of 500, the number of syntactic classes about 15.
2. The features themselves are often hierarchically organized and can be generalized by dropping the most specific features. For example, the word "*mass*" has *abnorm.physobj.finding.lesion* as its semantic features. The more general semantic class, *abnorm.physobj.finding* covers more words and *abnorm.physobj* covers an even larger set of words.

Word Valence

In addition to resonance between pairs of words, our linguistic model considers the *valence* of individual words. This is the preference of words for certain types of complements. A verb prefers to have one direct object and not two. A noun that is the object of a preposition prefers not to have a verbal complement.

We model the valence probability of a word in a given parse structure by the number and type of arcs into the word and from the word. Valence statistics are tabulated from the same set of annotated training sentences that are used for resonance statistics.

Parser Algorithm

The parser uses dynamic programming to calculate the highest probability dependency structure for a sentence. The probability of a given structure is the product of the resonance probability of each arc in the dependency graph, and of the valence probability of each word in the sentence. This is summarized in the following formula:

$$\Pr(\textit{dependency structure} \mid \textit{sentence}) = \prod \Pr(A \square B \mid A, B, \textit{direction}, \textit{distance}) \cdot \prod \Pr(\textit{valence } A \mid \textit{arcs into } A, \textit{arcs from } A)$$

The first term in this formula is the product of the probability of each arc from A to B. This is conditioned on the features of A and of B, the direction of the arc (right or left) and the relative distance between A and B. The second term is the product of the valence of each word, conditioned on the combination of arcs into that word and arc from the word.

Semantic Interpreter

The syntactic parser plays an important role in normalizing a wide variation in sentence structure found in free text narrative. It is the following step, the semantic interpreter, that creates the structured record from the parser output. The dependency graph that the parser produces has unlabeled arcs between words to show modifier relations. The semantic interpreter applies rules based on semantic features to translate these arcs into the logical relations.

For example, the arc from "*large*" to "*mass*" in Figure 1 in interpreted as size of a finding, while the arc from "*increased*" to "*mass*" is interpreted as a clinical trend. These interpretations are based on the semantic features of the words and the direction of the arc between them in the surface structure parse. Sometimes a semantic interpreter rule must traverse more than one arc to identify a logical relation. The arcs between "*mass*" and "*within*" and between "*within*" and "*segments*" are interpreted as the location of a finding. Figure 2 lists the logical relations created from the dependency graph of Figure 1.

Size (modifies "mass", value = "large")
Number of (modifies "mass", computed value)
Location (modifies "mass", rel="within", value = "segments")
Direction (modifies "segments", value = "posterior")
Part of (modifies "segments", value = "lobe")
Direction (modifies "lobe", value = "right")
Direction (modifies "lobe", value = "upper")
Trend (modifies "mass", value = "increased")
Property (modifies "increased", value = "size")

Fig. 2 - *Logical relations that interpret arcs in the parse for the sentence in Fig. 1.*

The semantic interpreter bundles logical relations together into output frames that list attributes of a finding, of a therapeutic or diagnostic procedure, or of an anatomic structure. Figure 3 shows a frame rooted on the finding "mass" from the logical relations in Figure 3. Each line in this frame either is an attribute of the finding or is a refinement of the previous line in the frame (shown here by indentation).

Finding	=	*"mass"*
Size	=	*"large"*
Number of	=	1
Location	*"within"*	*"segments"*
Direction	=	*"posterior"*
Part of	=	*"lobe"*
Direction	=	*"right"*
Direction	=	*"upper"*
Trend	=	*"increased"*
Property	=	*"size"*

Fig. 3 - *Structured representation as a frame*

The semantic interpreter rules are derived from a set of hand-tagged training sentences. The system developer uses a graphical interface to indicate the logical relations associated with a training sentence. The system builds rules by comparing the parse graph for a training sentence with a target logical relation to find a proposed rule that creates that logical relation. Some semantic features of the training example may not be essential to the rule. If the training sentence has "mass", the rule does not necessarily need to require the semantic feature "lesion", but might generalize to cover all words with the feature "finding". Generalizations of a proposed rule are tested on all other training sentences and the generalization is selected that covers as many training sentences as possible without making errors. Before a rule is finally accepted, the human system developer is given a chance to edit it.

Two processing steps remain before a frame is turned into the final output of the system, structured database entries. Information about a finding is often spread across multiple sentences. The system needs to merge together frames that relate to the same object. In addition, the terms in a frame must be mapped into a controlled vocabulary. These steps in our system have not yet been implemented.

RESULTS

A prototype system has been developed and trained on a corpus of thoracic radiology reports. Current evaluation efforts have focused on software validation and determining whether system specifications are appropriate (proof of concept). End-user testing has yet to proceed.

A preliminary evaluation of our parser algorithm was performed following the methodology used in the DARPA-sponsored Message Understanding Conferences [17]. The parser was evaluated using a ten-fold cross-validation study design. Given the preliminary state of the parser design, we did not perform a formal evaluation using an independent (i.e., unbiased) set of investigators. The evaluation described here was used to get a rough estimate on the amount of training required and accuracy that could be achieved by our design. The evaluation procedure is summarized as follows:

1. Ninety-two reports comprising 1115 sentences were randomly selected from a large pool of thoracic radiology reports. Each report was assigned to one of ten partitions.
2. For each sentence, one of the system developers hand-tagged all the attachments that an ideal parser would make. This served as the gold standard. A system developer performed this task because there are no universal specifications for how a parser should behave. (Note that the parser output is not an end-user result. A more formal evaluation using an independent set of investigators is planned after the current proof-of-concept evaluation).
3. The ten-fold cross validation study was performed. Each partition was used for testing. When the system was tested on sentences from one partition, the other nine partitions were used to train the system. Training involved learning resonance conditions between words. Evaluation of the test partition involved one of the developers scoring whether an arc in the output parse was: a) correct (i.e., agreed with the gold standard), b) missing, c) incorrect (connecting wrong set of words).
4. The performance of the parser is summarized according to the measures *recall* and *precision*. Recall is the percentage of correct arcs the parser identifies. Precision is the percentage of arcs reported that are correct. Table 1 shows the results as a function of resonance threshold value.

Resonance Threshold	Recall (%)	Precision (%)
0.0	90.0	89.4
0.3	88.4	91.0
0.5	82.5	93.4
0.7	77.7	95.2
0.9	62.7	97.7

Table 1 - *Parser performance from ten-fold cross-verification study using 1100 sentences.*

DISCUSSION

We present a statistical natural language processor based on resonance probabilities between word pairs. The parser uses no hand-coded rules, but rather gathers word affinity knowledge from training sentences whose dependency diagrams are manually specified. This ability to acquire knowledge is important for adaptability to new domains and writing styles. In the ten-fold cross validation study, the parser achieved high performance from a surprisingly small amount of training data. Recall and precision reached a percentile in the mid 80's from a little over one hundred training sentences and reached recall 90% at precision 89% by one thousand training sentences. The statistical models of resonance allow the system to generalize well, and behave gracefully in the presence of unseen grammar patterns.

Work is underway to improve the following aspects of the system: 1) Co-reference resolution, 2) dynamic modification of word features, 3) integration of existing electronic medical lexicons, 4) improved handling of conjunctive lists and parenthetical phrases, 5) handling of unknown words, 6) mapping system output representation to a controlled reference terminology such as SNOMED-RT.

REFERENCES

[1] P. Spyns. "Natural language processing in medicine: An overview," *Meth Inform Med* **35**:285-301, 1996.

[2] C. Friedman, P. Alderson, J. Austin, J. Cimino, S. Johnson. "A general natural-language text processor for clinical radiology," *J Am Med Informatics Assoc* **1**:161-174, 1994.

[3] P.J. Haug, D.L. Ranum, P.R. Frederick. "Computerized extraction of coded findings from free-text radiologic reports," *Radiology* **174**:543-548, 1990.

[4] N. Sager, M. Lyman, N.T. Nhan, L. Tick. "Medical language processing: Applications to patient data representation and automatic encoding," *Methods of Information in Medicine* **34**:140-146, 1995.

[5] M.B. Do Amaral, Y. Satomura. "Structuring medical information into a language-independent database," *Medical Informatics* **19**(3):269-282, 1994.

[6] A.M. Rassinoux, R. Baud, J.R. Scherrer. "Proximity processing of medical text," In: R O'Moore, S Bengtsson, J Bryant, J Bryden, eds. *MIE 90*. Springer-Verlag, pp. 625-30, 1990.

[7] G. Hripcsak, C Friedman, et. al. "Unlocking clinical data from narrative reports: A study of natural language processing," *Annals of Internal Medicine* **122**:681-688, 1995.

[8] C. Friedman. "Towards a comprehensive medical language processing system: methods and issues," *Proc. of the AMIA Fall Symposium*, pp 595-599, 1997.

[9] C. Manning, Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

[10] E. Charniak. *Statistical Language Learning*. MIT Press, Cambridge, MA, 1993.

[11] J. Allen. "Statistical Methods," *Natural Language Understanding*. Addison-Wesley Publishing, 195-204, 1995.

[12] A.L. Berger, S.A. Della Pietra, V.J. Della Pietra. "A maximum entropy approach to natural language processing," *Computational Linguistics*, **22**(1):39-71, 1996.

[13] D. Biber, S. Conrad, R. Reppen. *Corpus Linguistics: Investigating Structure and Use*. Cambridge University Press, 1998.

[14] D. Yuret. *Discovery of Linguistic Relations Using Lexical Attraction*. Ph.D. dissertation, Dept of Electrical Engineering and Computer Science, MIT, 1998.

[15] A. Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Ph.D. dissertation, Dept. of Computer and Information Science, University of Pennsylvania, 1998.

[16] D.B Johnson, R.K. Taira, A.F. Cardenas, D.R. Aberle. "Extracting Information from Free Text Radiology Reports," *International Journal of Digital Libraries* **1** (3):297-308, 1997.

[17] N. Chinchor, L. Hirschman, and D. Lewis. "Evaluate message understanding systems: An analysis of the third message understanding conference," *Computational Linguistics* **19**(3):409-451, 1993.