Automatic Segmentation of Clinical Texts

Emilia Apostolova, David S. Channin MD*, Dina Demner-Fushman MD, PhD∓,
Jacob Furst PhD, Steven Lytinen PhD, Daniela Raicu PhD

College of Computing and Digital Media, DePaul University, Chicago, IL 60604
*Northwestern University Medical School, Department of Radiology, Chicago, IL 60611
∓Communications Engineering Branch, National Library of Medicine, Bethesda, MD 20894

emilia.aposto@gmail.com, dchannin@nmh.org, ddemner@mail.nih.gov,
jfurst@cdm.depaul.edu, lytinen@cs.depaul.edu, dstan@cs.depaul.edu,

Abstract— **Clinical narratives, such as radiology and pathology reports, are commonly available in electronic form. However, they are also commonly entered and stored as free text. Knowledge of the structure of clinical narratives is necessary for enhancing the productivity of healthcare departments and facilitating research. This study attempts to automatically segment medical reports into semantic sections. Our goal is to develop a robust and scalable medical report segmentation system requiring minimum user input for efficient retrieval and extraction of information from free-text clinical narratives. Hand-crafted rules were used to automatically identify a high-confidence training set. This automatically created training dataset was later used to develop metrics and an algorithm that determines the semantic structure of the medical reports.**

A word-vector cosine similarity metric combined with several heuristics was used to classify each report sentence into one of several pre-defined semantic sections. This baseline algorithm achieved 79% accuracy. A Support Vector Machine (SVM) classifier trained on additional formatting and contextual features was able to achieve 90% accuracy. Plans for future work include developing a configurable system that could accommodate various medical report formatting and content standards.

I. Background

Clinical narratives, such as radiology and pathology reports, are a growing electronically available source of information. Clinical texts are commonly dictated and transcribed by a person or speech recognition software, or are directly entered in text form by physicians. Even though efforts have been dedicated towards promoting clinical data entry in structured format ([1], [2], [3]), clinical data is most commonly entered in the form of free text, probably because of time constraints that require fast data entry and uninhibited expression power. However, data available in structured format is necessary for the purposes of research, quality assessment, interoperability, and integrated decision support systems. As a result, there is a growing need for Natural Language Processing (NLP) to automatically convert clinical free texts to structured formats. Using NLP could bring all the benefits of a structured database while not incurring the cost of structured data entry.

Numerous applications require processing of information present in clinical text. In major hospitals, physician need to retrieve documents from large clinical text repositories, for example for the purposes of case finding. Clinical text has also been used to identify patients that could benefit from paticipation in a study and are eligible for recruitment ([4], [5]), in surveillance (such as monitoring disease outbreaks) ([6], [7]), or for discovery of disease-drug associations ([8]) and disease-findings ([9]) associations.

The development of a clinical text information extraction (IE) or information retrieval (IR) approach starts with identification of the types of information present in clinical narratives. Each type of clinical text serves a particular clinical purpose that imposes a semantic template on the information present in the text. A radiology report, for example, is a clinical text that is a primary means of communication between a radiologist and the referring physician. Even though radiology report formatting standards vary across hospitals, imaging modalities, radiologists, and change with time, the nature of the report requires at a minimum the following types of information: description of the procedure, patient demographics and history, image findings and observations, usually accompanied by a conclusion. These distinct types of information are usually demarcated by appropriate formatting to facilitate the interpretation of the radiology report by a human reader.

Knowledge of the structure of radiology reports is a necessary pre-processing step for a number of IR and IE tasks. For example, the presence of a disease or abnormality in the patient history section should be treated separately from evidence of a disease or abnormality in the report findings for the purpose of accurate case retrieval. An IE system searching for the negation of a disease, needs to differentiate between negations describing the reason for the exam (e.g. *rule out pneumonia*) and actual report findings (e.g. *increased opacity in the right lower lobe could represent an early acute pneumonic process*).

II. Task Definition and Dataset

The goal of this research is automatic structuring of clinical texts into pre-defined sections that will serve as a pre-processing step to clinical text IR and IE tasks. The dataset consists of 215,000 free-text radiology reports selected randomly from 3 million reports over a period of 9 years and

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

TABLE I
RADIOLOGY REPORT SECTIONS.

Section Name	Description
1. Demographics	Header information including Patient Name, Age, Date of Exam, Accession Number.
2. History	Clinical history and reason for the exam.
3. Comparison	Comparison with previous studies, if available.
4. Technique	Exam procedure.
5. Findings	The observations and findings of the report.
6. Impression	Conclusion and diagnosis.
7. Recommendation	Recommendations for additional studies and follow up.
8. Sign off	Attending radiologist, transcriptionist, and date on which the report was signed off.

TABLE II
SAMPLE RULES USED TO IDENTIFY FINDINGS SECTION, EXPRESSED AS REGULAR EXPRESSIONS.

$\hat{}(finding|observation|discussion)s?:$

A case-insensitive application of this regular expression will match beginning of a line, followed by the header strings (optionally in plural form) and a colon.

$\hat{}(\backslash W*)(finding|observation|discussion)s?(\backslash W*)\$$

A case-insensitive application of this regular expression will match a line containing the header strings (optionally in plural form) optionally surrounded by non-alphanumeric characters.

representing 24 different types of diagnostic procedures. The majority of the reports were transcribed by a human typist, with a small portion transcribed by a speech recognition system.

Sections of interest were identified by examining the dataset and consulting relevant guidelines. The American College of Radiology proposed a guideline for communication of diagnostic imaging findings[10] recommending the following components of a radiology report: *demographics, relevant clinical information, procedures and materials, findings, potential limitations, clinical issues, comparison studies, impression, diagnosis, follow-up or recommendation, any significant patient reaction*. One hundred randomly selected reports from the dataset were used for preliminary data analysis and eight report sections were identified (Table I).

Loose text formatting is commonly used to structure the reports. In some cases sections are designated with appropriate headings. For example, the History section could be marked by a heading such as *Clinical history, History, Indications*; similarly the Findings section could be marked as *Findings, Observations, Discussion*. Transitions to new sections could be indicated by one or more blank lines, ASCII visual markers such as *** or - - -, or a change of case (Impression was often distinguished from Findings by all capital case). Often, related sections, such as Comparison, Procedure, or Findings appear together in one paragraph.

Our task is to automatically segment the text in radiology reports into sections corresponding to the eight types of information present in the report.

III. METHOD AND RESULTS

The preliminary data analysis revealed common, local formatting patterns that could be used to locate section headers and boundary markers. A rule based algorithm was developed to identify sections based on boundary markers with the intention of automatically creating a suitable training set. This training set was later used to develop a high-accuracy algorithm capable of segmenting into sections all 215,000 reports in the dataset.

A. Training Dataset

Two to three rules were developed to match headings for each of the eight sections of interest. For example, section *History* was identified by locating text between known History headings (such as *History:, Indications:*, etc) and another known heading identifying a different section. Table II lists sample rules used to identify the Findings section. A report is considered automatically segmented only if all sections of interest were identified by the hand-crafted rules. Even though only a small portion of all reports contain all sections of interest, the algorithm requires successful identification of all eight sections. This guarantees that section patterns not captured by the hand-crafted rules will not cause inconsistencies in the automatically created training set. The algorithm was applied to all 215,000 reports (minus the reports set aside for preliminary analysis and testing) and 3,000 reports (less than 2%) containing all 8 sections of interest following the hand-crafted patterns were identified and automatically segmented into sections. An independent randomly selected test set of an additional 200 reports was manually annotated.

B. Similarity Metric Algorithm

The segmentation task was modeled as a classification task involving assigning each report sentence to one of eight categories. The similarity of the sentence to training sentences belonging to each section was used as a metric. Since section headings and report content in general tend to consist of specialized and mostly standard vocabulary, a relatively simple sentence similarity metric was used to measure the distance from each sentence to the eight categories [11]. Sections from the 3,000 training reports were used to compute weight word vectors corresponding to the 8 sections of interest. Data was first pre-processed and sentence tokens designating dates and numbers were converted to a common pattern. The Gate Open Source NLP framework was used to annotate date named entities [12]. The set of all 2-word sequences (bi-grams) across the training reports was used to

TABLE III
RESULTS FROM CLASSIFYING SENTENCES FROM 200 RADIOLOGY REPORTS INTO ONE OF 8 PRE-DEFINED SECTIONS.

Section	Accuracy	Hits	Misses	Total Number of Sentences
Demographics	0.99	1273	9	1282
History	0.67	77	38	115
Comparison	0.78	43	12	55
Technique	0.35	47	87	134
Findings	0.56	501	395	896
Impression	0.40	120	182	302
Recommendation	0.22	7	25	32
Sign-off	0.94	970	61	1031
Total	0.79	3038	809	3847

compute vectors corresponding to the frequency of the bi-grams in text from each section. The counts were normalized using a common weight factor: $tf * idf$ (term frequency - inverse document frequency [13]). $Tf * idf$ increases the importance of a term proportionally to the number of times it appears in the document, but offsets it by the overall frequency of the word in the set of documents (corpus). A normalized bi-gram vector also was computed for each of the test sentences and the vector cosine distance to each of the 8 section word count vectors was measured. The algorithm annotates reports by processing each sentence sequentially. The hand-crafted rules for determining section headers used for preparing the training set are applied first. If the sentence matches one of the expected header patterns, the sentence section is identified. When a sentence does not follow a hand-crafted pattern (which is the norm), the sentence is assigned to the closest section measured in cosine distance. If the difference between distances is insignificant (based on an empirically determined threshold), the algorithm assigns the sentence to the section of the previous sentence. Table III shows the result from this base-line version of the algorithm.

C. Learning with Support Vector Machines

Next we developed a classifier that uses additional context and formatting features. Boundary and formatting features are necessary for distinguishing semantically related sections. For example, the Impression (or Conclusion) section is often a summary of the Findings section, and could be distinguished by a human reader only by means of formatting (Impression is often capitalized). Reports were analyzed and an appropriate set of features together with their corresponding set of permissible values were identified. Table IV summarizes computed sentence features that were applied to training a sentence classifier. We used Support Vector Machines (SVM) [14] as a classification technique. SVM is a state-of-the-art classification technique proven to perform well on related NLP tasks and is a logical first choice. The classifier was trained on the features of the sentence, and on the features of surrounding sentences, using a sliding window of the previous and next report sentences. The segmentation task was modeled, in the same way as in the baseline approach, as a text classification task assigning each sentence to one of eight predefined sections. SVM classifiers were trained for each of the eight categories using sentence and surrounding sentence features. The eight classifiers were combined via one-vs-all classification - the category of the classifier with the largest output value was selected. Table V summarizes the results of applying the classifiers (trained on the automatically generated training set) on the test set of 200 reports.

TABLE IV
SENTENCE FEATURES USED FOR TRAINING A CLASSIFIER.

Sentence Orthography	Possible orthographic types are *All Capitals*, *Mixed Case*, or presence of a *Header pattern*, such as a phrase at the beginning of a line followed by a colon.
Previous Sentence Boundary	Formatting boundary separating the current and previous text sentences. Possible values are white space containing new lines, white space without new lines, non-alphabetic characters, or the beginning of the file.
Following Sentence Boundary	Formatting boundary separating the current and next text sentences. Possible values are white space containing new lines, white space without new lines, non-alphabetic characters, or the end of the file.
Cosine Vector Distance	Distance from the current sentence to each of the eight sections' word vectors.
Exact Header Match	This feature specifies if the sentence contains a header identified as belonging to one of the sections in the training data.

Formatting and boundary features significantly improved the classification of semantically related sections such as Findings and Impression. The Recommendation section proved to be hardest to classify as sentences expressing recommendations are often interleaved with sentences from the Impression and Findings section.

IV. CONCLUSION AND FUTURE WORK

Ad-hoc solutions to the problem of clinical text segmentation have been proposed in the past. As medical report formatting and standards vary across hospitals, a solution relying on a pre-determined training set is not practical.

TABLE V
SVM CLASSIFICATION RESULTS.

Section	Accuracy
Demographics	0.99
History	0.87
Comparison	0.86
Technique	0.92
Findings	0.91
Impression	0.89
Recommendation	0.78
Sign-off	0.99
Total	0.90

Similarly, solutions relying on the existence of an annotated training set also pose a practical problem as manually annotating a large corpus of medical reports is a resource consuming effort.

We suggest a two-phase algorithm. In the first phase, domain knowledge is used to identify report header rules for the automatic creation of a high-confidence training set. In the second phase, the automatically created training corpus is used to train a classifier that assigns a section heading to each sentence of a medical report. Our goal is to develop a scalable and robust medical report segmentation system that could be applied in large hospital settings. This study establishes that existing NLP techniques could be successfully applied to solving the report segmentation problem. We were able to achieve an accuracy of 79% using a baseline rule-based algorithm and an accuracy of 90% using an SVM classifier.

Future work involves developing a configurable system that could be used for various medical report formatting and hospital standards. The system will allow users to define semantic sections and a set of rules (expressed as regular expressions) to be used to automatically create a high-confidence training set for a given report type. After creating the training set, users will be able to specify formatting features to be used to train a classifier. This will provide the flexibility needed to process clinical texts from various sources. Our end goal is to facilitate information retrieval, extraction and data mining of clinical narratives by automating report segmentation and contributing the developed NLP modules to initiatives such as the Open Health Natural Language Processing Consortium [15].

REFERENCES

[1] A. van Ginneken and M. Verkoijen, "A Multi-Disciplinary Approach to a User Interface for Structured Data Entry," *STUDIES IN HEALTH TECHNOLOGY AND INFORMATICS*, pp. 693–697, 2001.

[2] N. Cheung, V. Fung, Y. Chow, and Y. Tung, "Structured Data Entry of Clinical Information for Documentation and Data Collection," *STUDIES IN HEALTH TECHNOLOGY AND INFORMATICS*, pp. 609–613, 2001.

[3] S. Rosenbloom, R. Miller, K. Johnson, P. Elkin, and S. Brown, "Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems," *Journal of the American Medical Informatics Association*, vol. 13, no. 3, pp. 277–288, 2006.

[4] S. Pakhomov, J. Buntrock, and C. Chute, "Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier," *Journal of Biomedical Informatics*, vol. 38, no. 2, pp. 145–153, 2005.

[5] M. Electronic, "Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure," *Am J Manag Care*, vol. 13, no. part 1, pp. 281–288, 2007.

[6] W. Chapman, L. Christensen, M. Wagner, P. Haug, O. Ivanov, J. Dowling, and R. Olszewski, "Classifying free-text triage chief complaints into syndromic categories with natural language processing," *Artificial Intelligence in Medicine*, vol. 33, no. 1, pp. 31–40, 2005.

[7] J. Haas, E. Mendonca, B. Ross, C. Friedman, and E. Larson, "Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients," *AJIC: American Journal of Infection Control*, vol. 33, no. 8, pp. 439–443, 2005.

[8] E. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman, "Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 87–98, 2008.

[9] H. Cao, M. Markatou, G. Melton, M. Chiang, and G. Hripcsak, "Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics," in *AMIA Annual Symposium Proceedings*, vol. 2005. American Medical Informatics Association, 2005, p. 106.

[10] ACR, "Acr practice guideline for communication of diagnostic imaging findings," 2005. [Online]. Available: http://www.acr.org/guidelines

[11] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[12] D. Cunningham, D. Maynard, D. Bontcheva, and M. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," 2002.

[13] G. Salton and M. McGill, *Introduction to modern information retrieval*. McGraw-Hill, Inc. New York, NY, USA, 1986.

[14] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2000.

[15] "Open health natural language processing (ohnlp) consortium," 2009. [Online]. Available: http://www.ohnlp.org