Automatic Segmentation of Clinical Texts - Preliminary Results

Emilia Apostolova, David S. Channin MD*, Dina Demner-Fushman MD, PhD$^{\mp}$,
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
$^{\mp}$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, and knowledge of their structure is necessary for enhancing the productivity of the healthcare departments and facilitating research. This paper presents a preliminary study attempting 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 the purpose of facilitating retrievability and information extraction from free-text clinical narratives. Hand-crafted rules are used to automatically identify high-confidence training reports, which are later used to develop a metric that quantifies the semantic structure of the reports. A word-vector cosine similarity metric combined with several heuristics is used to classify a report sentence into one of several pre-defined semantic sections. This algorithm achieves an accuracy of 79%. Plans for future work include developing a classifier that uses additional text formatting and section boundary features, as well as a limited surrounding context and we expect that we will be able to achieve an accuracy close to human annotator performance.

1 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 a speech recognition software, or are directly entered in text form by physicians. Even though effort has 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 the use of Natural Language Processing (NLP) for the purpose of automatically converting clinical free texts to structured formats.

The processing of the information present in clinical texts has numerous applications. Starting with the needs of hospital physicians and the sheer vastness of clinical text repositories in major hospitals, the most obvious use is relevant document retrieval, for example for the purposes of case finding. Clinical texts have also been used to identify patients that could benefit from a study for subject recruitment ([4], [5]), for the purposes of surveillance such as monitoring disease outbreaks ([6], [7]), or for the discovery of disease-drug associations ([8]) or disease-findings ([9]) associations.

The formulation of clinical text information extraction (IE) or information retrieval (IR) task would naturally start with identifying what types of information are 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. The radiology report, for example, is a clinical text that serves the purpose of a primary means of communication between the 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 minimum the following types of information - description of procedure and patient demographics and history, image findings and observations, usually accompanied by a conclusion. These distinct types of information are usually accompanied 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, presence of a disease or abnormality in patient history should be treated separately than 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*).

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.

Figure 1: Radiology report sections.

2 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 collected over a period of 9 years and describing 24 different types of diagnostic procedures. The reports were transcribed via a speech recognition software or a human typist.

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 lim-*

itations, clinical issues, comparison studies, impression, diagnosis, follow-up or recommendation, any significant patient reaction. A 100 randomly selected reports from the dataset were manually annotated for preliminary data analysis and 8 sections were identified (Figure 1).

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 tasks is to automatically segment the text in radiology reports into sections corresponding to the 8 types of information present in the report.

3 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. 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 2 lists sample rules used to identify Findings section.

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

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.

Figure 2: Sample rules used to identify Findings section, expressed as regular expressions.

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 the successful identification of all 8 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 test set) 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 additional 200 reports was manually annotated.

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 is used as metric. Since section headings and report content in general tend to consists 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. 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 [11]. The set of all 2-word sequences (bi-grams) across the training reports was used to 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). Tf-idf increases the importance of a word proportionally to the number of times it appears in the document, but offsets it by the overall frequency of the word in the 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 algorithms 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 empirically determined threshold), the algorithm assigns the sentence to the section of the previous sentence. Figure 3 shows the result from this base-line version of the algorithm.

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

Figure 3: Results from classifying sentences from 200 radiology reports into one of eight pre-defined sections.

4 Conclusion and Future Work

This algorithm is intended to serve as a baseline for developing a classifier that will use additional context and formatting features. Boundary and formatting features were not considered at this stage, and they 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). The results are good considering how limited are the features involved in classifying a sentence into a report section type. We expect that training a classifier based on the sentence context and a small number of additional sentence features will result in accuracy close to a human annotator. Table 1 summarizes computed sentence features to by applied in future work to training a sentence classifier. The classifier will be trained on the features of the sentence, and on the features of a surrounding sentences, using a sliding window of at minimum the previous and next report sentences.

Our goal is to develop a scalable and robust medical report segmentation system that could be applied in large hospital settings. The system requires minimum domain knowledge input specified as pattern rules, and does not depend on a manually annotated training set. This preliminary study establishes that existing NLP techniques could be successfully applied to solving the report segmentation problem. Our end goal is to facilitate future

Sentence Orthography	Possible orthographic types are *All Capitals*, *Camel Case*, or presence of a *Header pattern*, such as a phrase at the beginning of a line followed by a semicolon.
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.

Table 1: Sentence features intended to be used for training a classifier.

information retrieval, extraction and data mining of clinical narratives by automating report segmentation.

References

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

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

[3] S. Rosenbloom, R. Miller, K. Johnson, P. Elkin, S. Brown, Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems, Journal of the American Medical Informatics Association 13 (3) (2006) 277–288.

[4] S. Pakhomov, J. Buntrock, C. Chute, Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier, Journal of Biomedical Informatics 38 (2) (2005) 145–153.

[5] M. Electronic, Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure, Am J Manag Care 13 (part 1) (2007) 281–288.

[6] W. Chapman, L. Christensen, M. Wagner, P. Haug, O. Ivanov, J. Dowling, R. Olszewski, Classifying free-text triage chief complaints into syndromic categories with natural language processing, Artificial Intelligence in Medicine 33 (1) (2005) 31–40.

[7] J. Haas, E. Mendonca, B. Ross, C. Friedman, E. Larson, Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients, AJIC: American Journal of Infection Control 33 (8) (2005) 439–443.

[8] E. Chen, G. Hripcsak, H. Xu, M. Markatou, C. Friedman, Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study, Journal of the American Medical Informatics Association 15 (1) (2008) 87–98.

[9] H. Cao, M. Markatou, G. Melton, M. Chiang, 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). URL `http://www.acr.org/guidelines`

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