MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Automatic Generation of Repeated Patient Information for Tailoring Clinical Notes

Frank Meng, Ricky K. Taira, Alex A. T. Bui, Hooshang Kangarloo, Bernard M. Churchill

Medical Informatics Group, University of California, Los Angeles, Los Angeles, CA 90024

Abstract

Dictating clear, readable, and accurate clinical notes can be a time-consuming task for physicians. Clinical notes often contain information concerning the patient's medical history and current medical condition which is propagated from one clinical note to all follow-up clinical notes for the same patient. In this paper, we present a system which, given a clinical note, automatically determines what information should be repeated, and then generates this information for the physician for a new clinical note. We use semantic patterns for capturing the rhetorical category of sentences, which we show to be useful for determining whether the sentence should be repeated. Our system is shown to perform better than a baseline metric based on precision/recall results. Such a system would allow clinical notes to be more complete, timely, and accurate.

Keywords:

Natural language processing, Clinical Reporting Systems, Computerized Patient Records

Introduction

Within a series of clinical notes for a given patient, some of the patient information is often repeated, being carried over from one clinical note in the series to the next. Physicians often spend much time both determining what information needs to be repeated and re-dictating this repeated information when creating a follow-up clinical note for a given patient. This paper describes a methodology to automatically determine what information should be repeated in a new note, given as input a previous note for the same patient in the same series of notes. Our belief is that such a system can reduce the total amount of time needed to generate patient clinical notes and can also lead to more complete and accurate notes. Completeness is achieved because the system will always propagate to subsequent notes information designated by the physician to be repeated. The accuracy of the repeated information is dependent on the accuracy of the previous note(s) in the series and there is the ever-present danger of propagating erroneous information. With a user interface which clearly presents the repeated information to the physician, the content of each clinical note can be thoroughly reviewed.

Work in structured reporting [1, 2] has addressed the issue of reducing the amount of time for a physician to generate routine patient reports. Many of these systems utilize templates, which provide a general structure of the report and allow the physician to fill in the details, and macros, which allow the physician to use a type of shorthand to generate text for the report [3]. Though these systems can allow faster generation of reports, they are good only for generating patient information which does not change from patient to patient. Clinical notes often contain some amount of patient information shared among most patients, but often the repeated information depends on the particular patient's presentation and history.

The determination of repeated patient information is similar to generating a summary of the clinical note. Text summarization addresses the issues of automatically identifying and extracting the key information within a document [4, 5, 6]. Generating repeated patient information may be a less complex task than text summarization in that clinical notes often contain language which is more constrained and less variable.

Methods

Text Summarization

Work in text summarization research has taken several forms. Our task is most closely associated to text summarization *extraction* [7], which extracts segments of the original document to be used in that document's summary. This technique does not involve text generation, which is the case with summary *abstraction* [8]. We believe extraction for generating repeated patient information is more suitable than abstraction for the following reasons:

1. Physicians often use particular phrases to describe medical observations or events, and a system which automatically generates text may incorrectly convey vital information.
2. There is usually no need for the system to generate new text because the repeated information is often copied over to the next report by the physician (sometimes verbatim).

Our techniques draw from ongoing work being done in text summarization, but we have tailored some of the existing methodologies to better suit the needs of our problem domain.

Generating Repeated Patient Information

Figure 1 shows the overall functional diagram of the system we describe in this paper. The system requires a past clinical note as input and, using the information contained in it, generates the repeated patient information for the subsequent note.

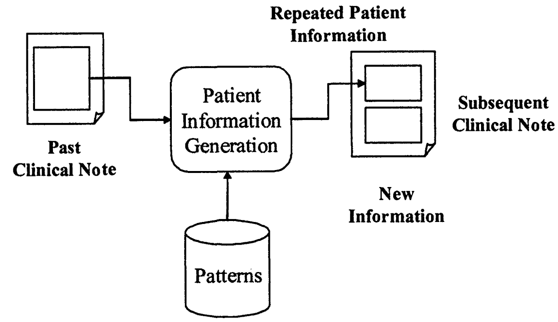

Figure 1 - System Diagram

Repeated patient information can contain standard patient information, such as the patient's name, age, or ethnicity, or it can contain information which may change from patient to patient, such as the patient's primary symptoms or illness or the patient's medical history. The following is an excerpt of a typical clinical note:

> *This patient has very profound general problems. He was born with the cord doubly wrapped around his neck just before birth. There was severe loss of fetal heart rate. He was delivered by cesarean section, but unfortunately has been profoundly retarded and quadriplegic since then. He is always in a wheelchair. He is unable to feed or dress himself or respond in any meaningful way. He has no language skills. He has had extensive trouble with gastroesophageal reflux and is now fed entirely by an gastrostomy tube. He also has had severe scoliosis for which there has been an anterior fusion. More recently, the problem has been one of hematuria and blood in the urine.*

Typically, physicians begin by indicating the major presentations of the patient. As the note progresses, the purpose of the current visit is mentioned as well as some other patient history. The following is an excerpt from the next note in the series:

> ***This patient had a severe problem at birth with the cord double wrapped around his neck when he was born. There was severe loss of fetal heart rate and he suffered irreversible ischemic cerebral damage.*** *He has significant problems with severe retardation.* ***He is quadriplegic. He is fed with a gastrostomy tube.*** *We have been having a number of urinary and fecal problems with this patient.*

The italicized portions show the patient information that was carried over from the first note.

Semantic Patterns

The basis for determining what information is repeated in our system is the *semantic pattern*. Semantic patterns have been used for information extraction [9, 10, 11, 12], paraphrasing [13], and other natural language processing applications. In our work, a semantic pattern is defined to be a sequence of semantic classes which has a one-to-one correspondence with a sequence of words and phrases in a clinical note. A semantic class is a categorization of a word or phrase which semantically identifies the concept the word or phrase represents. For instance, the word "disease" is represented by the semantic class "pName.abnormality". This means that it is a word that represents an abnormality in a patient. We use a natural language processor [14] which is capable of automatically assigning words and phrases to one of over 250 semantic classes. Figure 2 shows an example of a semantic pattern, illustrating its three components. Matching between two semantic patterns is defined to be a matching between their two associated semantic class sequences.

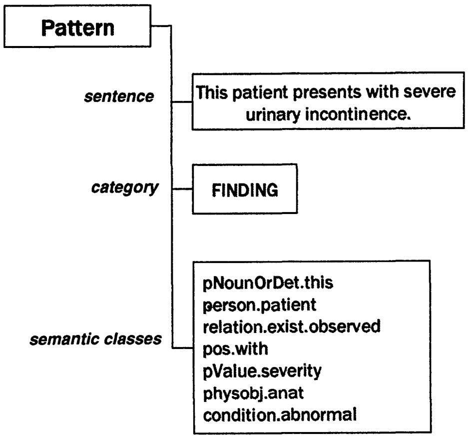

Figure 2 - Example Semantic Pattern

Categorizing Sentences using Semantic Patterns

The rhetorical status of sentences has been used successfully for text summarization [7], where each sentence is categorized based on its role in the document. We propose a similar approach to categorizing sentences in clinical notes. Whereas in [7], a Bayesian classifier [15] was used to automatically categorize sentences, our methodology uses semantic pattern matching. We have found that sentences of certain categories have a better chance of being repeated by the physician than sentences of other categories. For example, sentences which state a finding of a patient were found to be more readily repeated than sentences which reflect the physician's assessment of the patient's condition. To be more precise, for our corpus of urological clinical notes, we found that for sentences which state a urological finding of the patient, the repetition rate was 92%. It can be concluded that the rhetorical category of a sentence is a good indicator of whether that sentence will be repeated. We termed this particular set of sentences FINDING sentences, because they reflect findings of the patient by the physician. Examples of FINDING sentences would be, "the patient presents with meningomyelocele with a neurogenic bladder secondary to that" and "she is completely incontinent". Other sentence categories were also identified and these were used by the system to aid in differentiating FINDING sentences from other types of sentences. These categories include NORMAL (the patient has no abnormality), ASSESSMENT (the physician's assessment of the patient's situation), OBSERVATION (an observation made about the patient), TREATMENT (treatments the patient is undergoing or will undergo), DISCUSSION (communication with other physicians or family members), PSH (past surgical history), FMH (family medical history), EVALUATION (results of tests or examinations), and RECOMMENDATION (the physician's recommendation concerning the patient).

The major role that rhetorical categories play is to provide a clear definition of what exactly is repeated patient information. Our goal in the future is to provide a detailed set of rhetorical categories to the physician and allow the physician to determine, according to his/her needs, what type of information should and should not be included in repeated patient information. This would be done either implicitly, by training the system using a corpus of past clinical notes, or explicitly, by allowing the physician to tell the system which rhetorical categories should always be repeated.

Matching Semantic Patterns

Given an initial clinical note, the system uses a sentence boundary detector to decompose the note into its constituent sentences. Each sentence is matched against the patterns contained in the system's pattern set. The closest matching patterns will determine the sentence's categorization. The system's pattern set is generated beforehand using a tagged training corpus which will be discussed in more detail later.

The edit distance [16] has been widely used for approximate string comparison and is based upon the number of string edit operations needed to transform one string into the other. The edit distance metric is the least number of operations needed to do this transformation. Typically, string edit operations consist of inserting, deleting, or replacing characters. The edit distance was chosen over other metrics because it takes into consideration the order of the string, but it allows for approximate matches. Since order is important when dealing with semantic patterns in natural language, this metric is more suitable than other classification metrics such as naive Bayesian classifiers, which treat sequences of words as having no inherent order.

Because the constituent phrases of a given sentence could belong to several categories (e.g., a complex sentence composed of several simple phrases), we would like to match subsequences of the sentence against the patterns in the pattern set. We chose the Smith-Waterman algorithm [17], which is an edit-distance-related algorithm based on dynamic programming and is designed for approximate subsequence matching. The algorithm was found to perform well in determining whether a complex sentence contains a phrase which is similar to the phrase represented by the pattern it is being matched against. To optimize performance and also to focus the domain, a sentence is only matched against a pattern if the sentence contains all of the pattern's associated keywords. Keywords of a pattern are those words which form the basis of the pattern's meaning. These keywords are assigned during the tagging process (described in detail later). For matching a sentence with a pattern from the pattern set, a sequence of semantic classes is generated for the words in the sentence. The similarity score is then computed between the pattern's semantic class sequence and the sentence's semantic class sequence. Figure 3 shows an example of a sentence being matched against a pattern.

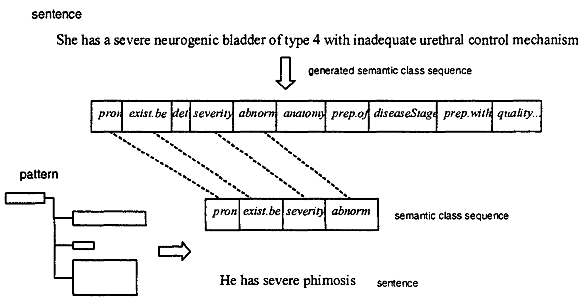

Figure 3 - Pattern matching

The highest scoring pattern when matched against a sentence using Smith-Waterman was used for determining the sentence's category. This is shown in the following equations, where *PS* is the pattern set, *s* is the sentence being matched, *cat(x)* is the rhetorical category of *x*, and $score_{SW}$ is the Smith-Waterman score:

$$P_{max} = p | max_{p \in PS}\{scoresw(p, s)\}$$

$$cat(s) = cat(p_{max})$$

Tagging the Corpus

The system's pattern set was generated using a tagged training corpus. Each sentence in the corpus was tagged by a human tagger to determine its rhetorical category. As mentioned, sentences often contain phrases belonging to several categories, so each sentence often was broken up by the tagger into several phrases and the individual phrase's category was tagged. An example of a tagged sentence with two tagged phrases is shown below:

[He also has had repeated urinary tract calculi] *FINDING:has:abnormal* which **[we have been treating]** *TREATMENT:treating* largely conservatively.

The above sentence, "He also has had repeated urinary tract calculi which we have been treating conservatively" is broken up into two phrases shown in bold. The tags of each phrase are shown in italics. The phrase "He also has had repeated urinary tract calculi" indicates the category FINDING and the phrase "we have been treating" indicates the category TREATMENT. In this example, the two phrases become the basis for two semantic patterns. Keywords are shown in the tag following the category name. The FINDING phrase, has associated with it the set of keywords "has" and "abnormal" and the TREATMENT phrase has "treating" as a keyword. The keyword "abnormal" refers to the semantic class of all abnormalities. A future direction of our work is to allow the system to automatically determine the salient keywords for each rhetorical category.

Results

For evaluation, the system's pattern set was generated based on tagged data described in the previous section. We generated

1060 patterns using approximately 650 tagged sentences. Our test corpus consisted of 162 sentences which mention urological abnormalities. For determining correctness, we tagged the test corpus sentences as to whether the sentence was actually repeated by the physician or not. The following rule of thumb was used: within a series of clinical notes for a particular patient, sentence A was deemed repeated by the physician if there exists another sentence B within any following clinical note which contains basically the same content as A. Exceptions to this rule include cases where A is mentioned for a particular examination, and B, containing the same basic content as A, is mentioned in the context of an entirely different examination. A and B, in this case, would not be considered a repeating pair of sentences.

We compared our system with a "Baseline" metric, which uses the position of a sentence in the note to determine the score for that sentence. Baseline assigns a score which is indirectly proportional to the sentence's distance from the beginning of the note, where the sentence closest to the beginning is assigned the highest score. The reasoning behind Baseline is that sentences that are located towards the beginning of the note are often included in the repeated patient information. We found that Baseline performs surprisingly well overall and was a good metric with which to compare our semantic pattern-based approach.

The results are presented in Comparison of Semantic Patterns and Baseline as a recall versus precision graph, where the X-axis represents the 11 standard levels of recall (0 – 100%) and the Y-axis represents the precision level. Our Semantic Patterns approach is shown as diamonds and Baseline is shown as squares. As can be seen, Semantic Patterns outperformed the Baseline overall, showing that semantic patterns are effective in capturing the features necessary to determine the FINDING rhetorical category of a sentence. We believe that even better results can be obtained given more training data.

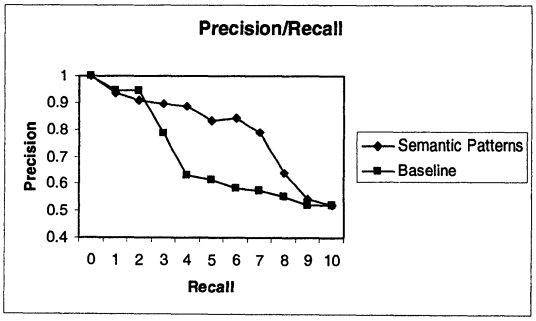

Figure 4 - Comparison of Semantic Patterns and Baseline

Discussion

The main contribution of this work is the use of a semantic pattern-based approach for automatically determining what patient information should be repeated in a clinical note. Semantic patterns were shown to be able to capture enough information to determine the rhetorical category of sentences contained in clinical notes. This is because semantic patterns representing the sentences of clinical notes reflect the way the author of the document intended to convey information concerning a certain topic or concept. We were able to show that physicians often repeat sentences of certain rhetorical categories, such as sentences which state facts of the patient's presentation. We used the Smith-Waterman algorithm for finding the similarity between a sentence and a pattern representing a particular category. This algorithm was chosen because of its capability to match subsequences within larger sequences. Since many of our patterns were phrase-based, this turned out to be a good choice.

Future work will consist mainly of improving the matching process between two semantic patterns. Work is currently being done to utilize the syntactic structure of language to prune out unnecessary phrases before a match score is calculated. This will help focus the similarity metric. There is also work currently underway which is focusing on determining semantic equivalence on the phrase level, which will allow the system to correlate semantic patterns that do not look alike on the surface but have the same underlying meaning. Determining the penalty for mismatches between subsequences is an issue that also needs to be addressed. We are looking into using a correlation metric such as mutual information to determine the crucial words of a particular rhetorical category.

Finally, repeated information in clinical notes is not limited to just propagating the same patient information from the first note to the last. Physicians often add new findings, test results, or procedures performed which become part of the patient's ongoing medical history. This new information either comes from a previous clinical note, or it may come from other source documents such as surgical, radiology, or laboratory reports. Our system must also be able to synthesize and summarize this new information and include this as new repeated patient information in subsequent clinical notes.

Conclusion

We presented a methodology for automatically generating repeated patient information in a series of clinical notes using natural language processing and semantic patterns. Semantic patterns are used to determine rhetorical categories for sentences in the clinical notes, and based on the rhetorical category, the system determines whether the sentence should be repeated in a subsequent note or not. Our system was trained on a corpus of urological clinical notes, and it was able to categorize sentences into several rhetorical categories better than a reasonable baseline metric based on a precision/recall measurement. We believe this methodology shows great potential in being able to reduce the amount of time and effort physicians must spend for generating and maintaining patient records.

Acknowledgements

This work was supported in part by a grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) PO1-EB00216, and from the National Library of Medicine, T15-LM07356.

References

[1] Langlotz CP and Meininger L. Enhancing the expressiveness and usability of structured image reporting rystems. In *Proceedings of the 2000 AMIA Annual Symposium*, 2000; pp. 467-471.

[2] Moorman P, Ginneken AM, Siersema PD, Van der Lei, J and Van Bemmel JH. Evaluation of reporting based on descriptional knowledge. *JAMIA*, 2:365-373, 1995.

[3] Gilbert P, and Light S. A simple methodology for developing clinical reporting macros. presented to SAS User Group International, 1996.

[4] Edmundson HP. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, Vol. 16, No. 2, p. 264-285, 1969.

[5] Radev DR, Hovy E and McKeown KR. Introduction to the special issue on summarization. *Computational Linguistics,* Vol. 28, No. 4, p. 399-408, December 2002.

[6] Hovy E, and Lin C. Automated text summarization in SUMMARIST. In Advances in Automatic Text Summarization, ed. Mani I and Maybury M, 1999.

[7] Teufel S, Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics,* Vol. 28, No. 4, p. 409-445, December 2002.

[8] Hongyan J, and McKeown KR. Cut and paste based text summarization. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL'00). May, 2000.

[9] Riloff E. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the 11th National Conference on Artificial Intelligence* (AAAI-93), p. 811-816, 1993.

[10] Riloff E. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (AAAI-96), p. 1044-1049, 1996.

[11] Muslea I. Extraction patterns for information extraction tasks: a survey. The AAAI-99 Workshop on Machine Learning for Information Extraction, 1999.

[12] Huffman SB. Learning information extraction patterns from examples. In *Proceedings of the 1995 IJCAI Workshop on New Approaches to Learning for Natural Language Processing,* ed. Wermter S, Sheler G, and Riloff E, 1995.

[13] Barzilay R and Lee L. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL*, p. 16-23, 2003.

[14] Taira RK, Soderland SG, and Jakobovits RM. Automatic structuring of radiology free-text reports. *Radiology* 2001; 21:237-245, 2001.

[15] Manning CD, and Schutze H. *Foundations of statistical natural language processing.* The MIT Press, Cambridge, MA, 1999.

[16] Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707-710, 1966.

[17] Smith TF and Waterman MS. Identification of common molecular subsequences. *Journal of Molecular Biology*, Vol. 147, p. 195-197, 1981.

Address for correspondence

UCLA Medical Informatics Group
924 Westwood Blvd.
Suite 420
Los Angeles, CA 90024; fmeng@itmedicine.net