Automatically extracting clinically useful sentences from UpToDate to support clinicians' information needs

Rashmi Mishra, DDS, MPH;[a] Guilherme Del Fiol, MD, PhD;[a]
Halil Kilicoglu, PhD;[b] Siddhartha Jonnalagadda, PhD;[c]
Marcelo Fiszman, MD, PhD[b]

[a] *Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA*
[b] *Lister Hill Center, National Library of Medicine, Bethesda, MD, USA*
[c] *Natural Language Processing Group, Mayo Clinic, Rochester, MN, USA*

Abstract

Clinicians raise several information needs in the course of care. Most of these needs can be met by online health knowledge resources such as UpToDate. However, finding relevant information in these resources often requires significant time and cognitive effort.

Objective: To design and assess algorithms for extracting from UpToDate the sentences that represent the most clinically useful information for patient care decision making.

Methods: We developed algorithms based on semantic predications extracted with SemRep, a semantic natural language processing parser. Two algorithms were compared against a gold standard composed of UpToDate sentences rated in terms of clinical usefulness.

Results: Clinically useful sentences were strongly correlated with predication frequency (correlation= 0.95). The two algorithms did not differ in terms of top ten precision (53% vs. 49%; p=0.06).

Conclusions: Semantic predications may serve as the basis for extracting clinically useful sentences. Future research is needed to improve the algorithms.

Introduction

Clinicians' patient care information needs are common and frequently unmet [1]. Most of these information needs can be met by online health knowledge resources like Medline and UpToDate [2]. However, clinically useful information is not always easy to find [3]. The most useful information for the care of a specific patient may be buried within long documents or fragmented across multiple documents and resources. Therefore, methods are needed to help clinicians identify clinically useful information efficiently and effectively.

Research on information extraction and summarization has been done in the biomedical text-mining domain, but most previous studies have been restricted to titles, abstracts, and metadata from Medline records [4-7]. More recently, the focus has shifted to extracting and summarizing information from the full-text of biomedical journals [8]. Although biomedical journals are sometimes useful for clinical decision making, they are not designed to directly answer clinicians' information needs [3]. On the other hand, resources such as UpToDate provide expert reviews on clinical topics with the goal of helping clinicians meet their patient care information needs. Although UpToDate documents provide summary recommendations on specific topics, these documents are still relatively long, often with over 200 sentences.

The overall goal of our research is to generate automatically knowledge summaries to support patient care decision making. Our approach consists of extracting clinically useful sentences from relevant documents using semantic natural language processing (NLP) methods. Specifically, in the present study we aimed at designing and assessing an algorithm that extracts clinically useful sentences on treatment recommendations for specific conditions from UpToDate documents.

Background

Clinicians' information needs. A seminal study by Covell et al. found that clinicians raise two questions out of every three patients seen and that 70% of these information needs go unmet [9]. A recent systematic review identified several studies that confirmed Covell's findings [1]. The review also identified significant barriers that limit clinicians' ability to meet their information needs, especially clinicians' lack of time and perception that an answer cannot be easily found in the available resources. In our research, we aim to address these barriers by reducing the time and cognitive effort that clinicians need to devote seeking for information.

Information extraction and summarization. Overall, text summarization can be classified into two types: 1) extractive summarization; and 2) abstractive summarization. In extractive summarization, the sentences are selected based on their relevance and key words. In abstractive summarization, novel sentences based on important concepts are created [8]. However, this method has many underlying challenges and is less popular than the extractive method.

Researchers have investigated both extractive and abstractive text summarization of the biomedical literature. Fiszman et al. designed a method that generates graphical abstractive summarization based on semantic interpretation of biomedical text [5]. Reeve et al. used the Unified Medical Language System (UMLS) to extract semantically related sentences for summaries [10]. Another method was proposed by Jin et al. to generate gene summaries from Medline abstracts based on the selection of information rich sentences [11]. Agarwal and Yu presented a method to extract figures in the biomedical literature based on a sentence classification system for selection of sentences from the full text [12]. Despite providing a foundation for our research, most prior studies have focused on assisting biomedical researchers, such as in generating new hypothesis. Unlike these studies, our goal is to summarize clinically useful recommendations to assist patient care decision making.

Previous Related Work. In a previous study, we assessed the feasibility of generating knowledge summaries composed of relevant sentences extracted from Medline citations [7]. The system consists of a pipeline that integrates multiple NLP tools and information retrieval resources, including the UMLS Metathesaurus [13] for concept extraction, SemRep for semantic predication extraction, [14] and MedRank for sentence ranking. The system achieved a high precision in extracting sentences related to the topic of interest, but MedRank did not perform well extracting the most clinically useful sentences. In the present study, we focused on full-text documents rather than abstracts and used a different approach to sentence ranking, which is described in the Method section.

SemRep. SemRep is a semantic NLP parser that uses underspecified syntactic analysis and structured domain knowledge from the UMLS [14]. SemRep extracts a set of semantic predications that consist of a subject (e.g., a medication), an object (e.g., a condition), and a predicate (e.g., 'TREATS'). Predications extracted by SemRep can be loaded into a relational database for further processing according to the needs of specific applications [15]. An example of a sentence and its SemRep output is listed below in Table 1. Our underlying assumption is that clinically useful treatment sentences generate a higher density of treatment-related predications than other sentences. This assumption served as the basis for designing our algorithm.

Method

The study method consisted of: 1) developing a gold standard composed of UpToDate sentences that were manually annotated regarding their clinical usefulness; 2) processing UpToDate documents with SemRep to generate sentence predications; 3) designing candidate algorithms to identify clinically useful sentences and selecting best candidate algorithms for the evaluation phase; and 4) comparing the performance of the selected algorithms.

Gold Standard. The gold standard consisted of a training set with 5 UpToDate treatment documents and a test set with 12 documents on the treatment of four conditions: coronary artery disease (CAD), hypertension (HT), depression, and heart failure (HF). The 12 documents consisted of the 3 most frequently accessed documents on the treatment of each of the 4 conditions according to UpToDate's usage log. In the gold standard, sentences were annotated according to a 5-point scale that rated the clinical usefulness of sentences. The scale was designed according to previous studies that showed clinician's preferences for patient-specific, objective, and actionable recommendations as opposed to study results and background information. Table 2 describes the rating instrument with examples.

The gold standard and the rating instrument were iteratively developed by three clinicians. First, one document was independently rated by two clinicians (RM, GDF), yielding an inter-rater agreement (linear weighted kappa) of 0.52. Disagreements were reconciled through consensus and the instrument was refined. In the next step a second document was rated independently by the same two clinicians (linear weighted kappa= 0.74) and was further refined. Next, another document was rated by RM and a third clinician who had not been previously exposed to the annotation instrument (linear weighted kappa=0.82). Given the high inter-rater reliability of the instrument, only one clinician (RM) rated the remaining documents.

Processing documents with SemRep. UpToDate documents in the training and test sets were obtained in XML format and then transformed by a script into SemRep's input format. The documents were then submitted to SemRep for batch processing. Last, the SemRep output was loaded into a relational database that was designed in previous research on Medline citations [15].

Designing candidate algorithms and selecting algorithms for final evaluation. Informed by documents in the training set, we designed several algorithm variations for preliminary analysis. The design was guided by manually inspecting sentences and their predications as well as by analyzing the frequency and types of predications generated by useful vs. not useful sentences. Candidate algorithms were then evaluated using the training set. Two algorithms that appeared to perform best were selected for the final evaluation: *Algorithm*1 and *Algorithm2*. Both algorithms were implemented as SQL statements that queried the predication database.

A*lgorithm1* was based on the density of predications in a sentence. The higher the number of predications generated by the sentence, the higher the sentence ranking. When two or more sentences had the same number of predications, the sentence that appeared later in the document received preference, since earlier sentences tended to be background sentences.

A*lgorithm*2 was similar to *Algorithm1*, except that it excluded from the final output sentences and predications that were considered to be less useful for clinical decision making. For this, the algorithm applied the following steps:

1) select predications with a predicate type of 'TREATS', 'ADMINISTERED', 'AFFECTS', 'PREVENTS', 'PROCESS_OF', 'compared_with', 'higher_than', 'lower_than', or 'same_as';

2) exclude sentences that contain one or more of the following predicate types: 'METHOD_OF', 'OCCURS_IN', 'COEXISTS_WITH', 'DISRUPTS', 'AUGMENTS', 'STIMULATES', 'INHIBITS', 'ASSOCIATED_WITH', 'CAUSES', 'LOCATION_OF', 'CAUSES', 'PART_OF', 'COMPLICATES', 'ISA', 'PRODUCES', 'PRECEDES', 'USES';

3) exclude sentences with predications whose subject is "placebo".

Evaluation. The two algorithms selected in the previous step were compared in terms of three outcome measures 1) top 10 precision (primary outcome); 2) average rating of the top 10 sentences; and 3) top 10 recall. Top 10 precision was obtained as the percentage of sentences among the top 10 ranked ones that were rated as Level 4 or 5 sentences in the gold standard. Average rating was obtained by calculating the average of the gold standard ratings for the top 10 sentences. Statistical significance was tested with Student's paired t-test for top 10 precision and recall, and Wilcoxon ranked sum test for the average rating.

Results

Documents in the training set had a total of 1293 sentences. Out of these, 743 (57.5%) sentences generated no predications. The average number of predications for sentences rated as Level 4 and 5 was 1.38 and 1.58 respectively. Other sentences had less than 1 predication on average.

Table 3 provides descriptive statistics for the test set. The 12 documents in the test set had 2833 sentences. Of these, 1623 (57.3%) sentences did not generate any predications. The correlation coefficient between sentence rating and average number of predications was 0.95. Sentences rated as Levels 4 and 5 generated 1.19 and 1.23 predications per sentence respectively, while other sentences generated less than 1 predication on average.

Table 4 presents the top 10 precision, top 10 recall, and average rating of the top 10 documents for both algorithms. No difference was found between *Algorithm1* and *Algorithm2* in terms of top 10 precision (53% vs. 49%; p=0.06) and average rating (3.5 vs. 3.4; p=0.4). *Algorithm2* was significantly better than *Algorithm1* in terms of top 10 recall (p=0.0002).

Discussion

In this study we aimed to develop and assess an algorithm that extracts clinically useful sentences from UpToDate. The ultimate goal is to automatically summarize treatment recommendations to help clinicians meet their patient care information needs. Both algorithms performed reasonably well but further studies are needed to improve the precision of extracted sentences. Using the algorithms designed in our study, about half of the sentences extracted by the algorithms in a knowledge summary would not be clinically useful. The two algorithms had equivalent performance in terms of the primary outcome (top 10 precision). Although *Algorithm2* had better top 10 recall, the absolute difference was only 7%. In addition, clinicians may favor precision over recall given the time constraints in busy clinical settings. However, clinicians' preference for precision over recall has not been studied and needs further investigation. In this study, *Algorithm1* may be a better option because it is simpler than *Algorithm2*. Perhaps most important is the finding that clinical usefulness is highly correlated with the number of predications in a sentence.

Although predication density seems to be a strong predictor for identifying clinically useful sentences, other approaches are needed to improve algorithm performance. We conducted an analysis of UpToDate sentences classified as useful vs. not useful in order to better understand their linguistic characteristics and to identify possible future directions. One characteristic that is significantly more prevalent in useful sentences is the use of *deontic modality* [16], particularly, of *obligative* type. This modality type is generally expressed by verbs such as *suggest* and *recommend*. In particular, when such verbs take as subject the first-person plural pronoun ("we"), the sentences that they appear in are generally useful sentences that indicate actionable statements, such as the following: "For patients with heart failure, we suggest amiodarone in preference to dofetilide."

Table 4. Performance of the algorithms on each document of the test set.

	Algorithm1			Algorithm2		
Document*	Top 10 precision	Average rating	Top 10 recall	Top 10 precision	Average rating	Top 10 recall
CAD1	50%	3.4	23%	30%	2.8	21%
CAD2	90%	3.9	21%	80%	3.9	31%
CAD3	20%	3.2	08%	20%	2.7	20%
Depression1	50%	3.4	11%	40%	3.4	12%
Depression2	40%	3.4	20%	30%	3.1	38%
Depression3	30%	3.1	23%	30%	3.1	33%
HF1	90%	3.9	15%	70%	4.0	19%
HF2	60%	3.7	08%	70%	3.8	14%
HF3	50%	3.5	09%	50%	3.2	17%
HT1	50%	3.4	19%	50%	3.4	31%
HT2	50%	3.4	41%	50%	3.4	50%
HT3	60%	3.5	42%	60%	3.5	46%
Average	53%	3.5	16%	49%	3.4	23%†

*CAD = Coronary artery disease; HT = Hypertension; HF = Heart Failure; †statistically significant

Another characteristic of useful sentences seems to be the high level of certainty, or lack of hedging. Hedging is often indicated by modal auxiliaries, such as *may* and *can*. In UpToDate, the use of hedging seems to be correlated with non-useful sentences. A highly speculative statement like the following would not be considered useful in clinical care. "For example, although the atrial myocardium may not be capable of sustaining AF in this setting, it may be able to generate and sustain atrial flutter."

Statements supported by specific, quantitative evidence are generally characterized as useful. In particular, it is noteworthy that all statements mentioning statistical significance with respect to some finding were deemed useful. On the other hand, statements indicating unspecific evidence are generally considered not useful. For example, in the following sentence hedging expressed with *may* also contributes to characterization of the sentence as not useful "Compared with MTX, there is less information available regarding the long-term safety of biologic DMARDs, and there is some evidence that the risk may be greater with these agents."

We also analyzed documents that yielded extreme precisions (high and low) to identify characteristics that may have contributed to creating these outliers. For example, the second document on coronary artery disease (90% top 10 precision) contained several evidence-based sentences such as "Angiotensin converting enzyme (ACE) inhibitors and angiotensin receptor blockers (ARBs) decrease cardiovascular mortality in post MI patients with systolic dysfunction and ACE inhibitors in most patients with an acute anterior MI." On the other hand, the third document on coronary artery disease focused primarily on describing the latest research on the treatment of this condition. For example, the document contained sentences like "A randomized trial comparing PCI with DES to minimally invasive direct coronary artery bypass surgery (MIDCAB) evaluated outcomes in 130 patients with isolated proximal LAD disease."

Limitations. This study had several limitations. First, although reliable the sentence usefulness scale was not clinically validated. For example, it is unknown whether sentences rated as more useful actually help clinicians' meet information needs. Second, the evaluation was limited to four conditions and treatment topics, and the study was limited to UpToDate. Thus, it is unknown whether the method can be generalized to other conditions, areas (e.g., diagnosis), and knowledge resources. Our algorithms used no information that is specific to UpToDate. Therefore, the algorithms are likely to generalize to other resources. Third, we did not test whether the retrieved sentences when combined produce a readable summary. Future studies are needed to design and assess summary presentation and readability.

Conclusion and Future work

Our study found that clinically useful sentences were strongly correlated with a higher number of predications. The two algorithms performed equally in terms of top 10 precision, achieving a reasonable top 10 precision (53% and 49%). Although usable, future research is needed to improve algorithm performance. Identifying deontic modality constructions and hedging seem to be promising approaches. Although SemRep is currently unable to identify these meta-predication constructions, work to implement this capability is underway. These topics have also been garnering much interest from the clinical and biomedical NLP communities recently [17, 18], and we plan to enhance our system based on insights from such research. Another potential approach is to employ machine learning sentence classification with SemRep predications as predictors.

Acknowledgments

This study was supported in part by grant number K01HS018352 from the Agency for Healthcare Research and Quality (AHRQ).

References

[1] G. Del Fiol, T. E. Workman, and P. N. Gorman, "Clinicians' Patient Care Information Needs: Preliminary Results of a Systematic Review of the Literature," *AMIA Annu Fall Symp,* 2012.

[2] S. E. Hauser, D. Demner-Fushman, J. L. Jacobs, S. M. Humphrey, G. Ford, and G. R. Thoma, "Using wireless handheld computers to seek information at the point of care: an evaluation by clinicians," *J Am Med Inform Assoc,* vol. 14, pp. 807-15, Nov-Dec 2007.

[3] J. W. Ely, J. A. Osheroff, M. L. Chambliss, M. H. Ebell, and M. E. Rosenbaum, "Answering physicians' clinical questions: obstacles and potential solutions," *J Am Med Inform Assoc,* vol. 12, pp. 217-24, Mar-Apr 2005.

[4] M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T. C. Rindflesch, "Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation," *J Biomed Inform,* vol. 42, pp. 801-13, Oct 2009.

[5] M. Fiszman, T. C. Rindflesch, and H. Kilicoglu, "Abstraction summarization for managing the biomedical research literature," presented at the Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, Boston, Massachusetts, 2004.

[6] J. R. Herskovic, T. Cohen, D. Subramanian, M. S. Iyengar, J. W. Smith, and E. V. Bernstam, "MEDRank: using graph-based concept ranking to index biomedical texts," *Int J Med Inform,* vol. 80, pp. 431-41, Jun 2011.

[7] S. R. Jonnalagadda, G. Del Fiol, R. Medlin, C. Weir, M. Fiszman, J. Mostafa*, et al.*, "Automatically extracting sentences from Medline citations to support clinicians' information needs," *J Am Med Inform Assoc,* Oct 25 2012.

[8] S. Bhattacharya, V. Ha-Thuc, and P. Srinivasan, "MeSH: a window into full text for document summarization," *Bioinformatics,* vol. 27, pp. i120-8, Jul 1 2011.

[9] D. G. Covell, G. C. Uman, and P. R. Manning, "Information needs in office practice: are they being met?," *Ann Intern Med,* vol. 103, pp. 596-9, Oct 1985.

[10] L. H. Reeve, H. Han, and A. D. Brooks, "The use of domain-specific concepts in biomedical text summarization," *Information Processing & Management,* vol. 43, pp. 1765-1776, 11// 2007.

[11] F. Jin, M. Huang, Z. Lu, and X. Zhu, "Towards automatic generation of gene summary," presented at the Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Boulder, Colorado, 2009.

[12] S. Agarwal and H. Yu, "FigSum: automatically generating structured text summaries for figures in biomedical literature," *AMIA Annu Symp Proc,* vol. 2009, pp. 6-10, 2009.

[13] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Res,* vol. 32, pp. D267-70, Jan 1 2004.

[14] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *J Biomed Inform,* vol. 36, pp. 462-77, Dec 2003.

[15] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, "SemMedDB: A PubMed-Scale Repository of Biomedical Semantic Predications," *Bioinformatics,* Oct 8 2012.

[16] F. R. Palmer, *Mood and modality*: Cambridge University Press, 2001.

[17] O. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *J Am Med Inform Assoc,* vol. 18, pp. 552-6, Sep-Oct 2011.

[18] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. i. Tsujii, "Overview of BioNLP Shared Task 2011," presented at the Proceedings of the BioNLP Shared Task 2011 Workshop, Portland, Oregon, 2011.