Dynamic Generation of a Table of Contents with Consumer-Friendly Labels

Trudi Miller[a], Gondy Leroy[a], Elizabeth Wood[b]
[a] School of Information Systems & Technology, Claremont Graduate University, Claremont, California, USA
[b] Lee Graff Medical & Scientific Library, City of Hope National Medical Center, Duarte, California, USA

Abstract

Consumers increasingly look to the Internet for health information, but available resources are too difficult for the majority to understand. Interactive tables of contents (TOC) can help consumers access health information by providing an easy to understand structure. Using natural language processing and the Unified Medical Language System (UMLS), we have automatically generated TOCs for consumer health information. The TOC are categorized according to consumer-friendly labels for the UMLS semantic types and semantic groups. Categorizing phrases by semantic types is significantly more correct and relevant. Greater correctness and relevance was achieved with documents that are difficult to read than with those at an easier reading level. Pruning TOCs to use categories that consumers favor further increases relevancy and correctness while reducing structural complexity.

INTRODUCTION

An increasing number of Americans use the Internet for health information. Conservative estimates place the number of health seeking Internet users at 40% of those adults with Internet access [1], while other estimates are closer to 80% [2]. There is a growing interest in online health information: fifty-two million Americans accessed health information online in 2000 [3], increasing to ninety-three million in 2003 [2]. Despite increased interest in online health information, many consumers are unable to understand the information they desire. A major obstacle for information seekers is the disparity between online health information's readability and their reading skill. Berland *et al.* [4] surveyed health information available online and found most to be accurate, but requiring at least a high school reading level to comprehend. The average literacy level for Americans is at the eighth or ninth grade level [5]. The National Assessment of Adult Literacy [6] found that ninety-three million Americans had either "below basic" or "basic" literacy level for prose. The gap between existing consumer health documents and the low national levels of health literacy is serious, as lower health literacy levels are related to increased hospitalization rates [7, 8].

It is unrealistic to expect that all consumer health information will be rewritten, or that the reading level of Americans will increase dramatically in the near future. Difficult health documents need to be transformed into a format that can be understood by those with minimal reading skill. Soergel *et al.* [9] advocate the use of an interpretive layer between health information generated by clinicians and its display to the consumer. The advantage of an intermediate layer is that it circumvents the monumental task of rewriting existing materials. An attempt to bridge the language of health professionals and consumers has led to the development of consumer health vocabularies [10]. Consumer health vocabularies provide mappings between health concepts expressed in expert terminology and the language used by the average consumer. However, they should adhere to three criteria according to Zeng *et al.* [11]: usefulness, clarity, and use of familiar words. An additional way to augment existing health information for consumers is to lead them to sections of interest. Consumers follow distinct searching patterns and patients tend to prefer terms related to diseases, syndromes, or body parts [12].

A table of contents (TOC) can provide both insight into the content and can be categorized around consumer interests. The most accessible narratives contain less background and present the most important content first [5, 13]. In a dynamic TOC the user chooses the categories that s/he is most interested in and can view its information immediately. We believe that the use of consumer-friendly categories within interactive TOC will help consumers with basic reading skills to access health information. For a TOC to be of assistance, it must be an accurate representation of the underlying text. This is why we are evaluating the correctness and relevancy of our TOC generating algorithm.

RESEARCH QUESTIONS

A TOC can assist consumers through organizing text into desirable categories. The most interesting categories can be viewed immediately, reducing the time taken to find relevant information. This reduces the volume of text to read, greatly assisting those with low reading skills. It is imperative that a TOC uses appropriate category headings and that text is correctly categorized under the headings. The Unified Medical Language System (UMLS) offers two levels of granularity in its Semantic Network: semantic types and semantic groups. The UMLS semantic groups were created by McCray *et al.* [14, 15] to reduce the complexity of the semantic types through aggregation. These groups take the 135 semantic types in the UMLS and classify them each into 15 more general groups. With both levels available, we evaluate whether semantic groups or semantic types can provide labels for a TOC that is created automatically for a document. It is also possible that easier documents do not use the type of clinical terminology that is present in the UMLS, so we also evaluate the effect of document readability.

METHODS

a. Categorization Using UMLS Semantic Network

TOCs generated using document metadata (like headers) require consistent markup or that the algorithm be modified for each different document source. Our algorithm is applicable to any consumer health information text and does not rely on manual labeling or document metadata. It uses medical domain knowledge encapsulated within the UMLS. Its Metathesaurus has over 5 million concepts already catalogued within the categories of its Semantic Network. This existing framework is robust and provides sufficient depth for consumer health documents.

b. Consumer Friendly Labels

Semantic types and groups have descriptive names that can be difficult for laypeople to interpret. 'Neoplastic Process' and 'Eicosanoid' are examples of difficult semantic types, while 'Physiology' is a semantic group that could be difficult to understand. Since these semantic categorizations form the foundation of the TOC generation algorithm, it is crucial that they be comprehensible to non-clinicians. Through consultation with a domain expert, consumer-friendly labels were created for each of the semantic types and groups. We provide the complete list of the consumer-friendly labels for the semantic types and groups at http://isl.cgu.edu/ConsumerHealth.htm. Later, we will use these understandable labels to solicit labels from consumers themselves.

b. Selection of Documents

Of the documents used in generating the TOCs, half had a difficult reading level and half had an easy reading level, calculated using Flesch's Reading Ease. Flesch's calculates a percentage between 1 and 100 for documents based upon the average sentence length and the number of syllables per word. It has comparable use in the literature [13, 16]. Chapman *et al.* [17] noted that readability measures are limited in evaluating complexity due to their focus on sentence and word length. We recognize this shortcoming and are concurrently developing an evaluation of document comprehensibility that considers vocabulary. Ten documents (score > 61) were categorized as easy. Ten documents (score < 50) were categorized as difficult. We generated TOCs based on semantic types and semantic groups for five of each readability condition.

d. Table of Contents Generation

Consumer health documents covering a variety of health topics were downloaded from the WebMD consumer health website. Noun phrases were extracted from each document using the General Architecture for Text Engineering (GATE) [18] software and its Noun Chunker, as shown in Figure 1A. Each phrase was searched for within the UMLS Metathesaurus (2005AB) using a customized stored procedure and all matching concepts (CUIs) were stored. If no matches were found for an entire noun phrase, then words were removed from the phrase until a match was found. If no match was found after reducing the phrase to a single word, then each individual word was searched for. If a phrase matched more than one concept, all matching concepts were used. If no match was found, the phrase was not included in the final TOC. The UMLS's Semantic Network was queried with all CUIs to retrieve all related semantic types and semantic groups, as shown in Figure 1B. If more than one semantic categorization matched a concept,

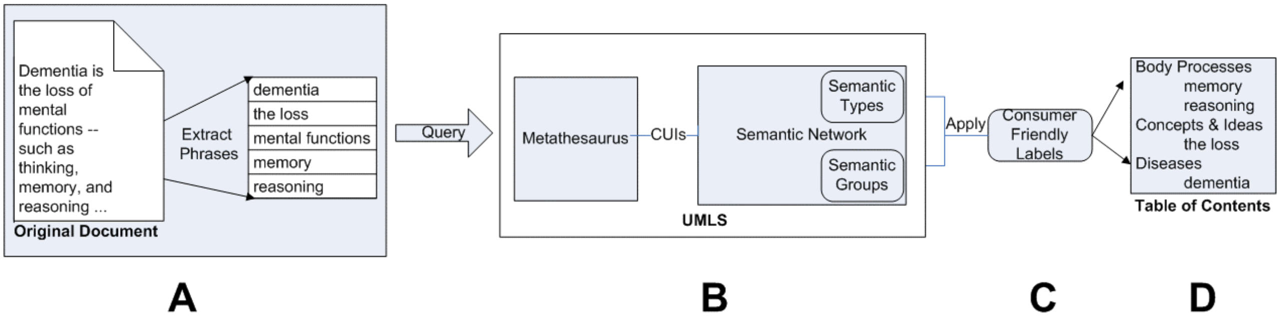

Figure 1 -- Overview of TOC generation algorithm architecture.

the concept was included under both categories in the final TOC. At this point, we evaluated all 135 semantic types and all 15 semantic groups. The semantic types and groups were replaced with our consumer friendly labels (Figure 1C). A TOC was generated from the phrase, its semantic type or group, and the phrase's original sentence (Figure 1D). A sentence could be assigned multiple headings. The following sentence could be assigned several categories: "Researchers continue to study drugs and other substances as possible treatments." It would be categorized as: Group Based on Job (from "Researchers"), Drugs (from "drugs"), and Treatments (from "possible treatments").

e. Expert Evaluation

A health information expert evaluated the TOCs from a consumer rather than a clinician perspective to ensure applicability to laypeople. Our expert has extensive experience with consumer health information, having set up consumer health information services and having taught medical terminology to those without a medical background. The expert considered two phrase level and one document level measures. The phrase level evaluation shows the correctness and relevancy of our approach in assigning phrases to the semantic groups and types. It also shows the usefulness of individual semantic types for a TOC. The relevancy of each phrase to the assigned semantic label was rated using a 7-point Likert scale where 1 represented "Strongly Relevant" and 7 represented "Strongly Irrelevant". A Likert scale was used to provide a gradient for relevance, allowing the expert to provide enumeration beyond relevant, irrelevant, and neutral. This allows for grading of phrases like 'painkiller' assigned to 'Chemical' to be graded lower than 'fluoride' but higher than 'wax'. The correctness of each phrase/semantic label assignment was assigned a value of "Yes" or "No". The document level measure provides a general assessment of the overall usefulness of the TOC. The overall TOC accuracy in reflecting the document content was rated using a 7-point Likert scale for the statement "This TOC accurately reflects the document content" where 1 represented "Strongly Agree" and 7 represented "Strongly Disagree".

RESULTS

a. Overview.

Twenty documents were evaluated in total. Ten documents were categorized using semantic types; five with easy readability scores and five with difficult reading scores. Ten documents were categorized using semantic groups; five with easy readability scores and five with difficult reading scores. We evaluated 4794 phrases. On average there were 239.7 phrases and 13 semantic group or 36 semantic type categories per document.

b. All Semantic Types and Groups.

Even though our algorithm did not act differently for different semantic types, correctness and relevancy scores vary substantially across the semantic groups based on the expert's evaluation, reaching the minimum and maximum possible values, 1 and 7 for relevancy and 0% and 100% for correctness (Table 1). The overall document accuracy was relatively low when all semantic types are used for a TOC, with both readability levels of semantic group documents measuring in the bottom half of the scale (Table 2). Categories like "Describes Amount of Space" and "Place" were consistently incorrect and irrelevant, while "Body Part" and "Disease" were correct and very relevant.

Semantic type categorization was significantly more relevant than semantic group ($p < 0.001$, Table 3), as were difficult documents compared to easy documents ($p < 0.001$, Table 3). There is a strong trend between semantic type categorization and increased correctness ($p < 0.10$, Table 3). Difficult documents were significantly more correct than easy ($p < 0.001$, Table 3). A significant interaction effect exists between semantic categorization type and difficulty level for correctness ($p < 0.001$, Table 3), indicating that difficult documents categorized by semantic type were more correct.

Table 1 -- Correctness (corr.) and relevance (relev.) for documents with high and low readability levels for semantic types and groups.

	Easy		Hard	
	Corr.	**Relev.**	**Corr.**	**Relev.**
Semantic Types				
Age Group	--	--	100	1.00
Animal	0	7.00	2	6.14
Area of the Body	52	1.79	17	4.60
Bacteria	100	1.00	100	1.00
Bird	0	4.00	--	--
Body Activity	20	5.00	18	5.50
Body Part	60	1.62	84	1.12
Body Substance	23	2.87	47	2.11
Cell	0	4.00	--	--
Cell Activity	100	1.00	--	--
Cell Part	0	7.00	0	7.00
Chemical	5	4.36	46	1.98
Chemical that Affects Living Things	5	5.54	43	1.76
Describes Amount of Space	< 1	6.24	3	6.12
Describes with Numbers	10	3.52	17	2.39
Describes with Words	88	1.12	59	1.31
Description of Medical Effect	25	3.20	39	2.25
Disease	76	1.32	72	1.36
Disorder	31	3.22	32	3.10
Drug	0	3.00	100	1.00

	Easy		Hard	
	Corr.	**Relev.**	**Corr.**	**Relev.**
Event	100	1.00	--	--
Family	100	1.00	--	--
Food	89	1.09	100	1.00
Gene	2	7.00	0	7.00
Governmental Action	0	7.00	0	7.00
Group Based on Job	84	1.00	35	1.64
Group of People with Things in Common	8	4.50	9	6.43
Grouping	0	7.00	14	7.00
Harmful Activity	18	2.75	27	3.56
How Cancer Grows	5	5.77	0	6.08
Idea	17	5.28	12	4.77
Injury or Poisoning	72	1.29	48	1.92
Job	10	6.00	0	5.25
Lab Test	0	7.00	0	4.61
Language	--	--	33	3.00
Living Activity	0	7.00	100	1.00
Living Being	--	--	17	6.00
Living Being Activity	22	2.85	33	3.00
Living Being Characteristic	100	1.00	33	3.00
Medical Activity	100	1.00	34	2.50
Medical Organization	0	6.44	2	5.89
Medical Thing	10	5.69	10	4.20
Natural Series of Events	0	7.00	17	6.00
Organization	--	--	0	6.00
Patients	--	--	100	1.00
Person	12	4.25	15	3.25
Place	< 1	6.89	2	6.49
Plant	9	4.38	40	1.67
Research Things	--	--	0	6.00
Result	12	4.91	8	4.62
Scientific Study	6	5.67	14	3.57
Series of Events	10	5.80	15	4.33
Substance	75	1.33	33	1.50
Symptom	77	1.18	67	1.33
Test	0	6.97	3	5.27
Thing	43	1.76	16	2.78
Things Groups Do	0	4.20	10	3.33
Things Used by Doctors or Dentists	1	4.25	6	5.33
Things You Do	47	2.11	25	1.60
Thinking	21	3.94	7	4.91
Time	42	1.68	14	3.14
Treatment	45	1.94	98	1.00
Virus	--	--	75	1.00
Semantic Groups				
Activities & Behaviors	31	2.60	32	3.09
Body Parts	78	1.26	35	2.18
Chemicals & Drugs	5	4.44	51	1.84
Concepts & Ideas	*10*	*6.23*	*17*	*5.64*
Tools	12	4.71	50	1.67
Diseases	38	2.32	56	1.67
Chemical Building Blocks	0	7.00	6	6.13
Places	2	6.79	1	6.69
Living Things	10	3.58	10	3.73
Things	9	2.06	33	2.04
Jobs	--	--	17	3.50
Groups	14	7.00	26	3.63
Events	3	6.08	4	6.00
Body Processes	21	3.31	16	3.95
Medical Procedures	5	6.36	18	3.60

Table 2 – Initial document accuracy results for semantic categorization and document readability (5 documents per condition).

	Semantic Group	Semantic Type	Totals
Easy	4.80	5.00	4.90
Difficult	4.20	2.91	3.56
Total	4.50	3.96	4.23

Table 3 – Semantic categorization and document readability means.

		Semantic Group	Semantic Type	Totals
Easy	Relev.	4.08	3.82	3.94
	Corr.	52	57	55
Difficult	Relev.	3.47	2.92	3.20
	Corr.	79	70	74
Total	Relev.	3.74	3.38	3.55
	Corr.	67	63	65

Table 4 – ANOVA results for relevancy and correct percentage results, run against semantic categorization and document readability.

Source of Variance	df	F	P
Relevancy			
Difficulty Level	1	99.06	.000
Semantic Grouping	1	27.91	.000
Interaction	1	3.65	.056
Error	4790		
Total	4794		
Correct %			
Difficulty Level	1	217.91	.000
Semantic Grouping	1	2.77	.096
Interaction	1	5.44	.000
Error	4790		
Total	4794		

c. Pruning Results.

Since our goal is to provide relevant TOCs for consumers, we evaluated our results a second time after removing the semantic groups considered irrelevant for consumers by our expert. This reduced the phrases to 163 per document and the categories to a mean of 9 semantic groups and 20 semantic types per document. We report only on relevancy and correctness for the phrase level analysis since our expert did not re-evaluate the pruned TOCs. However, by removing these irrelevant categories, a clearer picture of the impact of reading level and semantic categorization emerges.

Table 5 – Semantic categorization and document readability means after pruning.

		Semantic Group	Semantic Type	Totals
Easy	Relev.	3.18	2.49	2.82
	Corr.	52	73	63
Difficult	Relev.	2.82	1.86	2.19
	Corr.	79	87	83
Total	Relev.	2.80	2.16	2.48
	Corr.	67	80	73

Relevancy and correctness increase substantially through pruning, as expected. A significant relationship now exists between reading level and relevancy (F = 67.23, $p < 0.0001$).

DISCUSSION

Certain semantic categorizations were consistently incorrect and irrelevant in the initial categorization. Categories that do not meet relevancy criteria include those whose connection to medicine is peripheral, like ‘Bird’. Other categories are intangible, making it difficult to categorize, like ‘Group of People with Things in Common’ and ‘Concepts & Ideas’. Many types with high correctness and relevance are closely related to the medical field. For example, ‘Body Part’, ‘Virus’, ‘Treatment’, ‘Symptom’, ‘Disease’. ‘Body Parts’ and ‘Diseases’ are the most correct and relevant groups, and correspond closely with those topics that laypeople search with most frequently.

We suspect that difficult documents contain clinical language likely to be categorized correctly by the UMLS. These documents benefit more from a TOC, as their difficult content is harder for laypeople with lower reading levels to understand. With increased relevancy and correctness through pruning of TOC for difficult documents categorized by semantic type, we are optimistic that such an intermediary layer will afford consumers a great deal of benefit.

CONCLUSIONS

Labeling phrases with their semantic type has proven to provide higher overall accuracy, relevancy, and correctness than using the more general semantic group labels. Visualization of consumer health information through the generation of TOCs will continue, using the semantic types as the basis for labeling. The next phase of research is testing the TOCs with consumer groups.

ACKNOWLEDGMENTS

This work is supported by the NLM grant R21-LM008860-01.

REFERENCES

[1] L. Baker, T. Wagner, S. Singer, and M. K. Bundorf, "Use of the Internet and E-mail for Health Care Information," *Journal of the American Medical Association*, vol. 289, pp. 2400-2406, 2003.

[2] S. Fox and D. Fallows, "Internet Health Resources," Pew Internet & American Life Project, 2003.

[3] S. Fox and L. Rainie, "The Online Health Care Revolution," Pew Internet Organization, 2000.

[4] G. K. Berland et al., "Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish," *Journal of the American Medical Association*, vol. 285, pp. 2612-2621, 2001.

[5] C. C. Doak, L. G. Doak, and J. H. Root, *Teaching Patients with Low Literacy Skills*, 2nd ed. Philadelphia: J. B. Lippincott Company, 1996.

[6] S. White and S. Dillow, "Key Concepts and Features of the 2003 National Assessment of Adult Literacy (NCES 2006-471)." Washington, DC: U.S. Department of Education, National Center for Education Statistics, 2005.

[7] Institute of Medicine, "Health literacy: a prescription to end confusion," vol. 2006. Washington DC: National Academy Press, 2004.

[8] J. H. Root and S. Stableford, "Easy-to-Read Consumer Communications: A Missing Link in Medicaid Managed Care," *Journal of Health Politics, Policy and Law*, vol. 24, pp. 1-26, 1999.

[9] D. Soergel, T. Tse, and L. Slaughter, "Helping Healthcare Consumers Understand: An "Interpretive Layer" for Finding and Making Sense of Medical Information," presented at MEDINFO 2004, 2004.

[10] Q. T. Zeng and T. Tse, "Exploring and Developing Consumer Health Vocabularies," *Journal of the American Medical Informatics Association*, vol. 13, pp. 24-9, 2005.

[11] Q. T. Zeng, T. Tse, J. Crowell, G. Divita, L. Roth, and A. C. Browne, "Identifying Consumer-Friendly Display (CFD) Names for Health Concepts," presented at American Medical Informatics Association (AMIA) 2005, 2005.

[12] Q. Zeng, S. Kogan, N. Ash, and R. A. Greenes, "Patient and Clinician Vocabulary: How Different Are They?," presented at MEDINFO 2001, 2001.

[13] D. Gemoets, G. Rosemblat, T. Tse, and R. Logan, "Assessing Readability of Consumer Health Information: An Exploratory Study," presented at MEDINFO 2004, 2004.

[14] A. T. McCray, A. Burgun, and O. Bodenreider, "Aggregating UMLS Semantic Types for Reducing Conceptual Complexity," presented at MEDINFO 2001, 2001.

[15] O. Bodenreider and A. T. McCray, "Exploring semantic groups through visual approaches," *Journal of Biomedical Informatics*, vol. 36, pp. 414-432, 2003.

[16] C. A. Smith, P. Z. Stavrr, and W. W. Chapman, "In Their Own Words? A Terminological Analysis of E-mail to a Cancer Information Service," presented at AMIA 2002, 2002.

[17] W. W. Chapman, D. Aronsky, M. Fiszman, and P. Haug, "Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department," presented at AMIA Proceedings, 2000.

[18] Sheffield Natural Language Processing Group, "General Architecture for Text Engineering," 3.0 ed. Sheffield, UK: http://gate.ac.uk/, 2005.