Dynamic generation of a Health Topics Overview from consumer health information documents

Trudi Miller* and Gondy Leroy

School of Information Systems and Technology,
Claremont Graduate University,
130 E. Ninth Street, Claremont, CA 91711, USA
E-mail: trudi.miller@cgu.edu E-mail: gondy.leroy@cgu.edu
Website: http://ist.cgu.edu/leroyg
*Corresponding author

Abstract: Online health information use is increasing, but can be too difficult for consumers. We created a system that dynamically generates a health topics overview for consumer health web pages that organises the information into four consumer-preferred categories while displaying topic prevalence through visualisation. It accesses both a consumer health vocabulary and the Unified Medical Language System (UMLS). We evaluated its ability by calculating precision, recall, and F-score for phrase extraction and categorisation. We tested pages from three different consumer web sites. Overall, precision is 82%, recall is 75%, and F-score is 78%, and precision between sites did not significantly differ.

Keywords: consumer health informatics; information technology; information systems; natural language processing; text visualisation; Unified Medical Language System; UMLS.

Reference to this paper should be made as follows: Miller, T. and Leroy, G. (2008) 'Dynamic generation of a Health Topics Overview from consumer health information documents', *Int. J. Biomedical Engineering and Technology*, Vol. 1, No. 4, pp.395–414.

Biographical notes: Trudi Miller is a PhD candidate in the School of Information Systems and Technology, Claremont Graduate University. Her research interests include information systems, consumer health informatics, and user interface design. She has taught courses in systems analysis and design, programming, and information systems. She has published in the *Journal of the American Society for Information Science and Technology* and at several conferences.

Gondy Leroy is an Assistant Professor in the School of Information Systems and Technology at Claremont Graduate University. She received her PhD Degree in Management Information Systems from the University of Arizona in 2003. Her research interests are in text mining and intelligent interfaces for medical informatics and e-government. She has published in several journals such as *ACM Transactions on Information Systems*, *IEEE Transactions on Information Technology in Biomedicine*, and *Journal of Biomedical Informatics* among others. She sits on the editorial review board of the *Journal of Database Management* and several conferences. She received research funding from the National Library of Medicine, Microsoft, the National Science Foundation, and Edison International.

1 Introduction

The internet is frequently turned to as a source of health information by American consumers. Health searches are performed by 40–80% of all internet users (Baker et al., 2003; Fox and Fallows, 2003). Between 2000 and 2003, the number of Americans accessing health information online increased from 52 to 93 million people (Fox and Fallows, 2003; Fox and Rainie, 2000). This proliferation of searches shows rising interest in health information, but finding a relevant and accurate page is only the first part of learning about health topics from written materials. Once a consumer finds an appropriate web page, they must be able to read and understand it.

Most online health information requires at least a high school reading level to comprehend (Berland and Al, 2001). This is three years above the average American adult literacy level, which is at 8–9th grade (Doak et al., 1996). The National Assessment of Adult Literacy, based on a nationally representative sample of Americans over the age of 16, estimates that 93 million Americans have 'below basic' or 'basic' literacy for prose (White and Dillow, 2005). This means that they do not have the skills necessary to determine, e.g. which foods contain a specific vitamin by using reference materials. Many online documents require high reading levels due to their use of medical terminology, complex sentences, and passive voice. A system that reduces the amount of searching and reading required to find relevant sentences by grouping related health concepts may help bridge the gap between consumer's abilities and the reading levels of extant consumer health information.

We created a system that automatically creates a HTO from consumer health information. Our system uses natural language processing and existing biomedical resources to generate a content-based index of the underlying text organised into four categories: Body Parts, Diseases and Injuries, Drugs and Chemicals, and Medical Procedures. To create an accurate HTO, the system must accurately identify and group related health phrases, then categorise them into consumer-friendly categories. To do this, we have created a source-independent system leveraging existing natural language processing within our own custom algorithm. The HTO is displayed alongside the original document, visually indicates topic recurrence, and the user can select desired health topics to navigate to sections of interest. In this paper, we evaluate our system's output by evaluating precision, recall, and F-score for the automatically generated HTO for 27 consumer health web pages.

2 Theoretical foundation

With more patients searching for health information online, it is important they understand the pages that they find. Low health literacy has deleterious effects that extend far beyond the doctor's office.

2.1 Consequences of not understanding health information

The consequences of not understanding health information can negatively affect both a person's health and their utilisation of health care services. Lower health literacy has been linked to increased hospitalisation rates (Institute of Medicine, 2004; Root and Stableford, 1999) and, in general, poorer reported health (Rudd et al., 1999).

The American Medical Association (Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, 1999) found relationships between low health literacy, a decrease in understanding about the care they receive, and an increase in unwise health decisions. In contrast, increased health literacy makes a consumer more likely to engage in positive health behaviours (van Servellen et al., 2003) and leads to a feeling of empowerment (Fox and Fallows, 2003).

To a skilled reader it can be difficult to grasp the scale of the difficulties faced by consumers with low literacy levels. Williams et al. (1998a, 1998b) performed studies that measured patients' health literacy level and knowledge about their chronic disease. They found that the majority of asthma patients with third grade or lower health literacy levels did not understand that their puffer had specific instructions for use in order to receive an appropriate dose of medication (Williams et al., 1998a). Sixty percent of patients with inadequate health literacy levels and hypertension did not know that exercise lowers blood pressure (Williams et al., 1998b). The California Health Literacy Initiative found that 65% of participants with limited literacy skills avoided going to the doctor because of difficulties associated with completing paperwork (Bennett et al., 2003). Garbers and Chiasson (2004) studied immigrant Latina women aged 40 and older, and found those with inadequate functional health literacy were 16.7 times less likely to have ever had a Pap test, an important screening for cervical cancer. Low health literacy can also cause misunderstandings that are life-threatening. Williams et al. (1998b) found that almost two-third of diabetic patients with inadequate health literacy did not know that they should eat sugar when they became sweaty, shaky, or nervous. van Servellen et al. (2003) found that 66.6% of low-income, HIV-infected Latinos believed HIV could not be transmitted while on HIV medications. Consumers who do not understand their diseases are putting themselves, and others at risk through their confusion.

2.2 Current approaches to improve understanding

One way to make consumer health information easier to understand is to ensure its authors write it at an appropriate level. Current approaches to develop consumer-appropriate texts include manually establishing and enforcing writing guidelines for authors or automatic generation of texts targeted to consumer needs.

Writing guidelines to produce readable text help authors create documents that are understood by a wide range of consumers. The California Health Literacy Initiative provides resources for communicating health information in plain language and to English as a second language consumers (Literacyworks, 2007). The National Institutes of Health have created the Plain Language Initiative (National Institutes of Health, 2003), which requires the use of plain language in all new documents. Plain language is characterised by use of personal pronouns, active voice, lack of jargon, and design features such as bullets and tables. The US Department of Health and Human Services offers "Research-Based Web Design and Usability Guidelines", which outlines in great detail how to write and organise web-based content (Leavitt and Shneiderman, 2006). It makes recommendations such as using black font on a white background, using descriptive headings, highlighting critical data, and displaying items in a vertical list rather than contiguous text.

Writing guideline initiatives may assist authors in the future, but they do not address the vast body of consumer health information already written. There are thousands of

web pages, brochures, and prescription inserts already available. Rewriting all current materials according to these guidelines would require years of effort and could never overcome the mounting backlog as new information is written without using the guidelines. Ensuring that writers follow the guidelines is impossible, as there are multiple sets of guidelines and no policing agency for the internet.

In a perfect world, a consumer would go to a web site, answer a few questions, and in a few seconds a document written specifically for their needs and appropriate to their reading skill would appear in their web browser. Natural Language Generation (NLG) is the automatic creation of text that appears to be written by a human and has been tested in this regard. However, the algorithms that generate natural language are complex and currently require a lot of human-coded input information to be successful. Hirst et al. outlined the steps required for HealthDoc, an NLG system for health education documents (Hirst et al., 1997). Fundamental components of this system include a medical writer, the authoring tool for the medical writer, and the patient's medical record. None of these components are currently available to consumers using pre-written health information. They also require a lot of effort for the authoring agency, as well as specialised training.

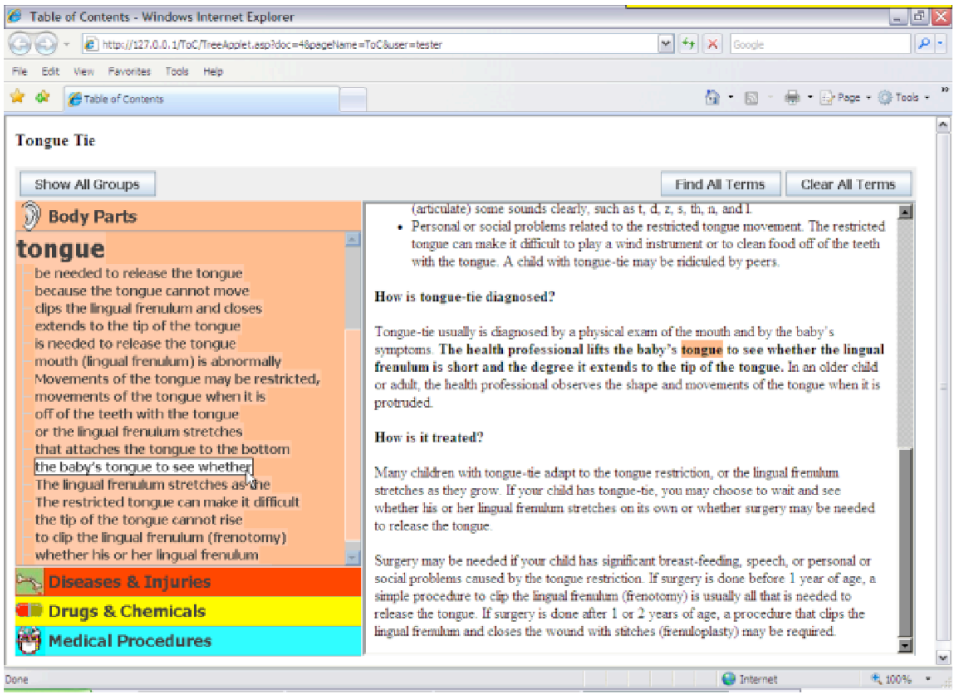
Even if one overlooks the large resource investment required for current NLG systems, their efficacy has come into question. Reiter et al. (2003) evaluated 2553 users of STOP, an NLG system that generated tailored letters aimed at quitting smoking. Subjects completed a questionnaire about their smoking behaviour and feelings about smoking. Unfortunately, subjects who received the STOP authored letters were no more likely to quit than those who had received a form letter. STOP's inability to increase the number of people who quit smoking shows that customisation alone does not ensure success.

3 Dynamic health topic overview to visualise consumer health information

Our system dynamically generates a HTO for consumer health information web pages and features consumer-preferred semantic categories as its headings. When displayed, the HTO appears next to the original document and can act as a visual overview or navigation aid (Figure 1).

Underneath each heading are related health phrases, reducing the number of phrases that the reader must process. Conceptual grouping causes semantically related phrases to be displayed under the same heading, such as 'congenital hypothyroidism' under the heading 'cretinism'. When the user selects a heading, all phrases below that heading are highlighted within the original document and their constituent sentences are bolded to make the information easier to find. This navigation aid reduces the amount of reading required, which may benefit weaker readers as they can focus on comprehending only the text that interests them. Manually assembling the data for a categorised health topic index would be as impossible as manually updating all extant health literature to adhere to writing guidelines. For the index to be plausible, it must be generated automatically. We discuss each processing step to build this HTO in detail below.

Figure 1 Consumer health topic index applet, with a single snippet selected. The related sentence is displayed in bold and the phrase is highlighted (see online version for colours)



3.1 Resources used

The system combines five existing natural language processing and lexical resources (see Table 1 for an overview), tailored each to meet the specific needs of this system, allowing the system to benefit from expert linguistic and medical informatics knowledge.

Table 1 Existing used by the HTO system

Resource name	*Description*	*Available online from*
General Architecture for Text Engineering (GATE)	Natural language processing software	http://gate.ac.uk
GATE-UMLS combined lexicon	271,157 phrases	http://isl.cgu.edu/ConsumerHealth.htm
SPECIALIST lexicon	330,345 items in 557,397 inflected forms	http://lexsrv3.nlm.nih.gov/SPECIALIST/
Open-Access Collaborative Consumer Health Vocabulary (Consumer Health Vocabulary)	156,826 phrases	http://consumerhealthvocab.org
Unified Medical Language System (UMLS)	4,335,846 phrases *(2007AA base installation with SNOMED CT)*	http://umlsinfo.nlm.nih.gov

3.1.1 General Architecture for Text Engineering (GATE)

Our system uses GATE (Sheffield Natural Language Processing Group, 2005; Cunningham et al., 2002) to perform natural language processing tasks that do not require special tuning for health text. GATE takes human-authored text, known as 'natural language', and breaks it into labelled parts that can be understood by computers for processing. Our system uses GATE's tokenizer, sentence splitter, Part-Of-Speech (POS) tagger, and noun phraser to perform the initial natural language processing.

GATE is not optimised specifically for biomedical applications, so the POS tagger used the GATE-UMLS combined lexicon (Leroy et al., 2006) instead of the standard version. It includes 271, 157 phrases and their POS tags optimised for biomedical text.

3.1.2 Open-access and collaborative consumer health vocabulary

The Open-Access and Collaborative Consumer Health Vocabulary (ConsumerHealthVocab.org, 2007) provides 156,826 consumer health phrases and their matching concepts within the UMLS. The consumer-friendly phrases are chosen through systematically reviewing candidates from MedlinePlus consumer queries (Zeng et al., 2005). It maps these consumer phrases to existing clinical concepts within the UMLS, which represents the language of clinicians.

3.1.3 Unified Medical Language System (UMLS)

The UMLS (National Library of Medicine, 2007) has three components: the Metathesaurus, the Semantic Network, and the SPECIALIST lexicon.

The Metathesaurus is a database comprised of 4,335,846 phrases grouped into related concepts (Table 1). It is comprised of several contributing source vocabularies, some parts of which are identified, through a field named 'term type', as exclusively for medical coding use or as obsolete terminology. Using the default installation, including the SNOMED CT vocabulary, those vocabularies not well suited to mapping against consumer text were removed from the database. Some concepts within the UMLS have been identified by the Consumer Health Vocabulary initiative (see previous section) as being irrelevant to consumer health language. These concepts have also been removed from our system's installation of the UMLS.

The Semantic Network assigns categories, known as semantic types, to all concepts in the Metathesaurus. The Semantic Network and Metathesaurus are often used to identify and semantically categorise medical phrases within medical text. For example, Elhadad et al. (2005) mapped phrases from clinical reports and studies to the UMLS to generate custom summaries. The UMLS was also used by MediClass to map medical concepts to identify vaccine reactions (Hazlehurst et al., 2005). Bashyam and Taira (2005) used a custom phrase chunker to isolate phrases, then mapped the phrases to the UMLS to create an anatomical index of the reports. Our consumer friendly categories are based upon semantic types. To select relevant semantic types, we looked to the health language used by consumers. The literature reports recurring patterns in the phrases used by consumers when searching health topics. McCray et al. (1999) looked up the semantic type groupings for 91,944 user queries from the NLM home page. They found that most belonged to four categories: Diseases and Pathologic Processes, Chemicals and Drugs, Procedures, or Anatomy. Similarly, Zeng et al. (2001) collected queries to the 'Find-A-Doctor' website of the Brigham and Women's Hospital in Boston. They found

the most popular categories to be Disease or Syndrome, Health Care Activity, Biomedical Occupation or Discipline, Body Part, Organ or Organ Component, and Finding. Tse and Soergel (2003) found that consumers most commonly use search phrases related to disorders, procedures, chemicals and drugs, and concepts and ideas. Using these findings, we show four consumer-friendly categories: Body Parts, Diseases and Injuries, Drugs and Chemicals, and Medical Procedures (Table 2). These are the four categories found by all to be important.

Table 2 Categories used in the HTO system compared with consumer-preferred term categories found in the literature

System categories	*McCray et al.*	*Zeng et al.*	*Tse and Soergel*
Body parts	Anatomy	Body part organ, organ or organ component	Chemicals and drugs
Diseases and injuries	Diseases and pathologic processes	Disease or syndrome finding	Disorders
Drugs and chemicals	Chemicals and drugs		
Medical procedures	Procedures	Health care activity	Procedures
Not in system categories		Biomedical occupation or discipline	Concepts and ideas

Body Parts, Diseases and Injuries, and Medical Procedures were found to be consumer friendly by all three of the studies. Drugs and Chemicals was found to be consumer friendly by McCray et al. (1999), and is also one of the major topic sections found in many consumer websites such as those used in this study.

In earlier work, we found that several certain semantic types are unsuitable for categorising health phrases for consumers (Miller et al., 2006). Due to their inaccuracy for consumer health information, vagueness, or lack of health relevancy, we omitted categories such as 'Concepts and Ideas' from the HTO.

An earlier, well-known effort to group the semantic types is McCray et al. (2001), who organised all 134 semantic types into 15 'Semantic Groups'. Our four HTO categories share similar grouping with these semantic groups (Table 3). However, we completed the groupings for our HTO based upon the consumer searching literature, described previously. After creating our categories, we compared them to the well-known semantic groups to determine whether our grouping of semantic types was similar to that of the semantic groups. We excluded semantic types such as "Molecular Biology Research Technique" from the semantic group 'Medical Procedures' because the concepts (e.g., yeast two hybrid system, two hybrid interaction trap) are not frequently the subject of searches and they do not logically fall under any of the four consumer-friendly headings. We excluded nine semantic groups entirely that were not found to be consumer preferred by more than one study (e.g., Activities and Behaviours, Objects).

The SPECIALIST lexicon is a biomedical and general English list that includes 330,345 medical terms and their 'base' forms (The Lexical Systems Group, 2007). The base form has a single value for all related lexical entries, including variations in spelling and plural forms. Conversion of a term to its base entry ensures that all similar words, irrespective of spelling, are represented by the same word. Our system uses this lexicon in conjunction with GATE to simplify the health phrases found within our text by using their base forms when available.

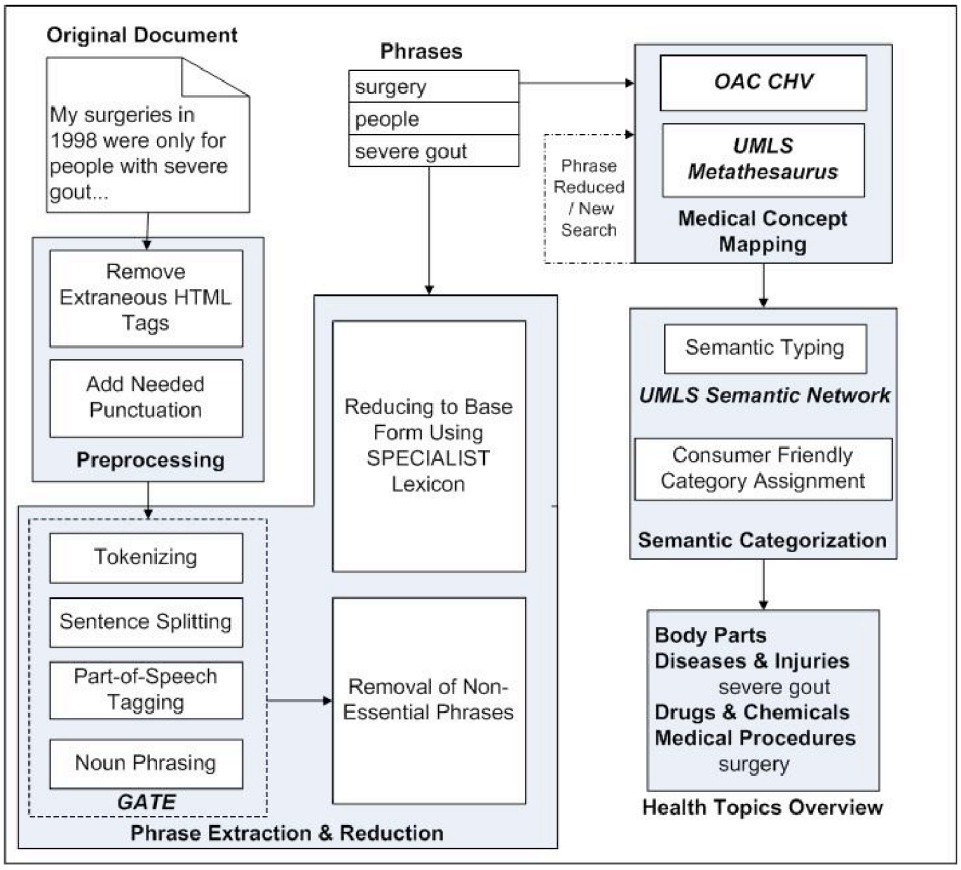
Table 3 Overview of the four primary HTO categories in terms of constituent semantic types and compared to semantic groups

HTO category	*Semantic type*	*Semantic group*
Body parts	Anatomical structure	Anatomy
	Body location or region	
	Body part, organ, or organ component	
	Body space or junction	
	Body system	
	Cell	
	Cell component	
	Embryonic structure	
	Fully formed anatomical structure	
	Tissue	
	Gene or genome	Genes and molecular sequences
Diseases and injuries	Acquired abnormality	Disorders
	Anatomical abnormality	
	Cell or molecular dysfunction	
	Congenital abnormality	
	Disease or syndrome	
	Experimental model of disease	
	Finding	
	Injury or poisoning	
	Mental or behavioural dysfunction	
	Neoplastic process	
	Pathologic function	
	Sign or symptom	
Drugs and chemicals	Antibiotic	Chemicals and drugs
	Biologically active substance	
	Chemical	
	Chemical viewed functionally	
	Enzyme	
	Hazardous or poisonous substance	
	Hormone	
	Immunologic factor	
	Indicator, reagent, or diagnostic aid	
	Neuroreactive substance or biogenic amine	
	Organic chemical	
	Pharmacologic substance	
	Receptor	
	Vitamin	
Medical procedures	Diagnostic procedure	Procedures
	Health care activity	
	Therapeutic or preventive procedure	

3.2 HTO mapping algorithm

Figure 2 shows our algorithm visually. The explanation below explains the details of each step. We use the sentence "My surgeries in 1998 were only for people with severe gout" as an example to illustrate each step.

Figure 2 Overview of the HTO algorithm (see online version for colours)



3.2.1 Original document

The document text is used to generate the HTO, independent of document markup such as the labelling of headers or titles. The system currently accepts HTML documents, because most web pages are in HTML format, but it could be expanded to accept additional formats. HTML is used as an umbrella term, as the system handles multiple tagging styles and formats such as XHTML and DHTML. When creating headings for the HTO, the algorithm only considers the document's vocabulary.

3.2.2 Pre-processing

The first step of the HTO system consists of pre-processing the web page by removing all unnecessary HTML tags: all custom HTML tags, stylesheets, and tags such as DIV, SUP,

and SPAN are removed. Periods are added to the end-of-tags, such as list elements (e.g., <\li>) and headings, to ensure their content is properly handled during sentence splitting (see below).

3.2.3 Phrase extraction and reduction

Tokenising. The first step is breaking the document into atomic words and punctuation, referred to as tokens. An example that has been tokenised is:

<T>*My*</T> <T>*surgeries*</T> <T>*in*</T> <T>*1998*</T> <T>*were*</T>
<T>*only*</T> <T>*for*</T> <T>*people*</T> <T>*with*</T> <T>*severe*</T>
<T>*gout*</T><T>.</T>

The blank spaces are also tokens, but have been omitted from this example for the sake of brevity. Our system uses the tokenizer included with GATE.

Sentence Splitting. Once all tokens are identified, the document can be broken into sentences. This is done based on punctuation. After sentence splitting, the example would appear as:

<Sen><T>*My*</T> <T>*surgeries*</T> <T>*in*</T> <T>*1998*</T>
<T>*were*</T> <T>*only*</T> <T>*for*</T> <T>*people*</T> <T>*with*</T>
<T>*severe*</T> <T>*gout*</T><T>.</T></Sen>

Part-of-Speech (POS) Tagging. Every word and punctuation within a document is a token. Words are assigned a POS tag that represents their lexical category (e.g., plural noun, adverb, verb past tense):

<Sen>**PRP$***My* **NNS***surgeries* **IN***in* **CD***1998* **VBD***were* **RB***only* **IN***for*
NNS*people* **IN***with* **JJ***severe* **NN***gout*.</Sen>

For specific details on the type of POS tagger used by GATE, see Hepple (2000). To augment the POS rules built into GATE, the GATE-UMLS combined lexicon was used. This provides additional and improved POS tag/word combinations that more accurately tag biomedical text than GATE can alone.

Noun Phrasing. Noun phrasing is the process of grouping together words, based on their parts of speech, to form a noun phrase:

<Sen><NP>*My surgeries*</NP> *in* <NP>*1998*</NP> *were only for*
<NP>*people*</NP> *with* <NP>*severe gout*</NP>.</Sen>

GATE uses rules that label noun phrases based on the sequence of the POS tags within a sentence. For example, the sequence '*severe gout*' is an adjective followed by a common singular or mass noun. This sequence of POS tags is recognised as a sequence representing a noun phrase, and is labelled as a noun phrase in the example above.

Removal of Non-Essential Phrases. Noun phrases sometimes begin with words that add nothing to the health meaning of the phrase. Examples include determiner pronouns such as 'both' and 'most', prepositions such as 'near' and 'before', and conjunctions such as 'or' and 'where'. These are all closed class words, which are words that belong to classes of words that do not typically allow new words to be added to them. The system removes

all closed class words from the front of the noun phrases. From the example sentence the noun phrase 'my surgeries' has had the closed class word 'my' removed:

<NP>*surgeries*</NP>; <NP>*1998*</NP>; <NP>*people*</NP>;
<NP>*severe gout*</NP>

Shortening the list of noun phrases benefits the user because it reduces the amount of reading as well as the length of time it takes to generate the HTO. If the entire phrase consists of a closed class word, a number, a single character, or in a list of stop words defined by the authors, it is omitted. In the example, this removes the phrase '1998', leaving us with:

<NP>*surgeries*</NP>; <NP>*people*</NP>; <NP>*severe gout*</NP>

The next step is reducing each phrase to its most basic form. For example, the phrase 'germs' can be reduced to the base of 'germ'. Each noun phrase is compared against the SPECIALIST lexicon and its base phrase is extracted. If no match is found within SPECIALIST, the original form of the noun phrase is used. Our system uses this approach rather than stemming to avoid reducing terms to the point of losing health semantic value and becoming unintelligible to consumers. In the example, 'surgeries' would be simplified to 'surgery', but the other phrases remain the same:

<NP>*surgery*</NP>; <NP>*people*</NP>; <NP>*severe gout*</NP>

3.2.4 Health concept mapping

Each noun phrase is searched for within the Consumer Health Vocabulary. If the phrase is found, all matching concepts are stored and searching stops. Phrases not matched within the Consumer Health Vocabulary are matched to the UMLS where possible. The system's copy of the UMLS Metathesaurus tool has been customised to improve accuracy and efficiency as described above. If one or more concept matches are found for a noun phrase, all concepts are stored and the search for that phrase stops. In the example, only one concept was found for each phrase:

surgery: 1 concept: C0543467

people: 1 concept: C0027361

severe gout: 1 concept: C0018099

If no match was found for the noun phrase, the algorithm attempts to identify a health concept for a sub-phrase. The first word within the phrase is removed if it is a modifier, such as an adjective, and the remaining noun phrase is used to search. This type of searching is referred to as head phrase matching (Riloff and Jones, 1999). For example, 'severe gout' is not found within either Consumer Health Vocabulary or the UMLS, so 'gout' is then searched for. If head phrase matching is unsuccessful, then each word within the phrase is examined individually, from left to right. If a match is found, the match proceeds on to the next step. If still no match is found, the noun phrase is omitted from the HTO.

3.2.5 Semantic categorisation

The semantic type is determined for each concept, using the UMLS Semantic Network. If multiple semantic types are found, all are stored. The Consumer Health Vocabulary provides a list of incorrect mappings, errors in semantic type identification from a consumer health perspective. When encountered, the system does not implement these erroneous mappings, based upon the Consumer Health Vocabulary recommendation. In the example, the system assigned semantic types as follows:

surgery: Therapeutic or Preventive Procedure

people: Population Group

severe gout: Disease or Syndrome

Once categorised within the Semantic Network, each noun phrase is also assigned a HTO category. All noun phrases that are not assigned to one of the HTO categories are dropped from the HTO. This results in the phrase 'people' being dropped from our example, as Population Group is not in any of our HTO categories:

surgery: Medical procedures

severe gout: Diseases and injuries

If a single noun phrase within a given sentence has multiple assignments of the same HTO category, the phrase is displayed only once.

3.2.6 HTO user interface

The final result is a list of noun phrases, concepts, and categories that are displayed in a Java applet (Figure 1). The four primary categories appear on the left side of the screen and the user can maximise each in turn if desired. Maximising the category means that it expands to show as many concepts as possible, minimising the other categories to their image and title. The size of the concept is relative to the number of noun phrases present within the document. More frequent terms are displayed in a larger font, making them stand out more. In Figure 1, the 'Body Parts' category is maximised and the concept 'tongue' has been expanded. When expanded, concepts will display part of the sentence, a 'snippet', for every noun phrase related to that concept in the document. When a sentence snippet is chosen, the text window on the right scrolls to the relevant sentence which is displayed in bold and the relevant phrase is highlighted.

4 Evaluation

To determine the algorithm's efficacy, an evaluation corpus was compiled from three popular consumer health websites: Family Doctor (FamilyDoctor.org), MedlinePlus (medlineplus.gov), and WebMD (webmd.com). MedlinePlus and WebMD were chosen because they are the two most popular consumer-focused sites according to Alexa, the Web Information Company (alexa.com). Family Doctor was chosen due to its award-winning site design that focuses on communicating with consumers. Nine web pages were downloaded from each of the three sites. Three topics were chosen from each of three categories common to all three sites: diseases, drugs, and treatments. Pages were chosen if they had more than 500 words and did not cover a topic used during algorithm

development and early testing. Topics included heart disease, acupuncture, drug information on Sitagliptin, and electroconvulsive therapy. We made sure none of the topics overlapped, providing the most generalisable findings by covering 27 different topics. All title images, navigation links, copyright information and other extraneous content were manually removed. We generated a gold standard by extracting all noun phrases and categorising them, where applicable, in either one of the four HTO categories. The gold standard was generated for the 27 documents before running the algorithm, ensuring objectivity for evaluation.

To evaluate the algorithm, we compared the phrases assigned to the categories in the HTO against the gold standard to calculate precision (1), recall (2), and F-measure (3).

$$\text{precision} = \frac{\text{correctly\_retrieved\_items}}{\text{all\_retrieved\_items}} \tag{1}$$

$$\text{recall} = \frac{\text{correctly\_retrieved\_items}}{\text{all\_correct\_items\_in\_document}} \tag{2}$$

$$F\text{-score} = \frac{2 \quad \text{precision} \quad \text{recall}}{\text{precision} + \text{recall}}. \tag{3}$$

An item is deemed to be correct only if both the noun phrase was accurately identified and the HTO classification matched the gold standard's classification. This means that the phrase 'kidney stones' classified as 'Disease and Injury' is correct, while 'kidney stones' classified as 'Drugs and Chemicals' is a false alarm. Precision represents the proportion of correct HTO items from all items retrieved. Recall represents the proportion of correct HTO items retrieved from all available correct items. The F-measure combines both precision and recall into one metric providing a balanced measure and ensuring that one metric is not being optimised at the expense of the other.

5 Results

5.1 Descriptive statistics

Each document within the corpus contained an average of 1166 words (Table 4). The average number of paragraphs per document was 46.

Table 4 Descriptive statistics for evaluation corpus

Source	*Words*	*Paragraphs*	*Noun phrases*
FamilyDoctor	8,695	328	2622
MedlinePlus	11,511	476	3613
WebMD	11,239	443	3619
Total	*31,485*	*1247*	*9854*

Leading closed class words were removed from 34.83% of all noun phrases. The average length of initial noun phrases was 10.78 characters, but noun phrases displayed within the HTO were reduced an average of 5.59 characters. Three types of unwanted noun phrases

were removed whose content was not health related or too vague to be of value. The three types are numeric, single character, and stop words (Table 5).

Table 5 Percent of noun phrases removed

Reason for removal	*Noun phrases* (%)
Stop words	16.57
Numeric	1.26
Single character	1.25
Total removed	*19.08*

Completely numeric noun phrases made up 1.26%. Single character noun phrases were 1.25% of the total. Stop words were determined during the development of the HTO to be either incorrectly classified or of little semantic value for consumers, including all closed class words as well as an additional 19 nouns such as 'problem'. Noun phrases consisting entirely of stop words comprised 16.57% of all noun phrases.

After removal of unwanted noun phrases, 97.31% of the remaining noun phrases were matched within the SPECIALIST lexicon. Of those phrases that were found within SPECIALIST, 28.95% were shortened, such as the original noun phrase 'surgeries' being reduced to 'surgery'. The remaining 68.36% of the noun phrases were already in base form.

The noun phrase was searched for within the Consumer Health Vocabulary first, and 83.59% of noun phrases were found and assigned one or more concepts, as shown in Table 6.

Table 6 Percent of noun phrases found within each medical vocabulary resource

Match found in	*Noun phrases* (%)
Consumer health vocabulary	83.59
UMLS (Custom)	8.67
UMLS	1.05
Not found	6.69
Total	*100*

Those not found in the Consumer Health Vocabulary were searched for in our reduced version of the UMLS. An additional 8.67% of noun phrases, such as 'Ibandronate sodium', were matched within our reduced version of the UMLS. An additional 1.05% of noun phrases found a match within the full UMLS, such as 'fungal'. These phrases from the full UMLS are not included in the HTO due to inaccuracy of the contributing vocabulary. The phrase 'fungal' is classified as a 'Functional Concept', which is not a semantic type displayed within the HTO. The decision to exclude the vocabularies was made during the system's design, prior to the experiment, as explained above. Noun phrases that did not match any of the sources made up 6.69%. Unmatched phrases included: 'The NCCAM Clearinghouse', 'Zephrex', and 'bookstore'.

Twenty-five semantic types contain 1% or more of all noun phrase entries (Table 7). Seven of the top nine semantic types account for 44.72% of all noun phrases, and are present within the HTO.

Table 7 Top UMLS semantic types

Semantic type	*Noun phrase allocation* (%)	*Present in HTO*
Pharmacologic substance	9.21	Y
Therapeutic or preventive procedure	6.22	Y
Disease or syndrome	5.56	Y
Sign or symptom	5.04	Y
Organic chemical	4.51	Y
Finding	4.38	Y
Professional or occupational group	3.46	
Body part, organ, or organ component	3.39	Y
Intellectual product	2.95	
Qualitative concept	2.76	
Quantitative concept	2.67	
Functional concept	2.37	
Temporal concept	2.18	
Manufactured object	2.03	
Idea or concept	1.96	
Spatial concept	1.96	
Population group	1.94	
Tissue	1.57	Y
Medical device	1.49	
Organism function	1.39	
Health care activity	1.32	Y
Laboratory procedure	1.20	
Amino acid, peptide, or protein	1.18	
Conceptual entity	1.17	
Diagnostic procedure	1.00	Y
All others (less than 1% allocated)	25.37	*Some*
No match	1.73	
Total	*100*	

Out of all noun phrases, 4,042 or 41.55% are categorised into one of the HTO categories: 630 in Body Parts; 1,702 in Drugs and Chemicals; 1,011 in Diseases and Injuries; 699 in Medical Procedures. This means that out of all noun phrases used, in spite of the 19.08% that were excluded as being unwanted, our system is still able to identify and classify 41.55% into consumer-preferred categories.

5.2 Test metrics

Compared against the gold standard, precision was 82%, recall was 75%, and the F-score was 78% across all documents in the evaluation corpus. The algorithm's performance was similar for all three types of websites: FamilyDoctor.org, MedlinePlus, and WebMD

(Table 8). Precision was around 80% (83%, 79%, and 84% respectively). Recall was almost as high, around 75% (76%, 77%, and 72% respectively). Due to the consistency of the precision and recall figures, the F-score was also between 75% and 80% (79%, 77%, and 78% respectively).

Table 8 Precision, recall, and F-score by document source and category

N = 27	*Precision* (%)	*Recall* (%)	*F-score* (%)
FamilyDoctor	83	76	79
MedlinePlus	79	77	77
WebMD	85	72	78
Overall	*82*	*75*	*78*
Body parts	75	86	78
Diseases and injuries	87	78	82
Drugs and chemicals	70	65	71
Medical procedures	50	77	54
Overall	*70*	*77*	*71*

The algorithm's performance across the HTO categories varied (Table 8). We found a main effect for category type for precision ($F(3, 102) = 8.82$, $p < 0.01$). Post-hoc contrasts showed that Body Parts and Diseases and Injuries were significantly more precise than Medical Procedures (Body Parts, $p < 0.01$; Diseases and Injuries, $p < 0.001$; Bonferroni adjusted).

We also found a main effect for category type for recall ($F(3, 99) = X$, $p < 0.01$). In this case, post-hoc contrasts showed that Body Parts had significantly higher recall than Drugs and Chemicals (Body Parts, $p < 0.001$; Bonferroni adjusted).

Finally, we also found a main effect for category type for the *F*-Score ($F(3, 98) = 11.32$, $p < 0.001$). Post hoc-comparisons showed that the *F*-Score for Body Parts, Diseases and Injuries, and Drugs and Chemicals were significantly higher than that of Medical Procedures (Body Parts, $p < 0.001$; Diseases and Injuries, $p < 0.001$; Drugs and Chemicals, $p < 0.02$; Bonferroni adjusted).

6 Discussion

Overall, the HTO system performed well, especially for phrases categorised as Diseases and Injuries and Body Parts. The elimination of unwanted phrases was able to substantially reduce the number of noun phrases the system must categorise and that the end user needs to see. The system also showed that it works equally well regardless of the document source. This holds great promise for its generalisability to other sites and authors in the future.

The performance of our HTO algorithm reflected some traditional problems common to natural language processing. Imperfect noun chunking led to omissions, such as 'heart attack or stroke' being chunked into a single noun phrase instead of two. As expected, word sense disambiguation also posed a problem. Some health conditions have abbreviations that are also common phrases: 'tips' can refer to a "transjugular

intrahepatic portosystemic shunt procedure" or the more common 'piece of advice'. This led to difficulty in determining which HTO category a phrase should fall under. In some cases, diagnosis refers to a disease, while others use it to refer to the process a physician completes in ascertaining the problem.

Two categories showed low performance. Upon evaluating the source of the errors, we were able to identify three problems responsible. The low recall rate of the Drugs and Chemicals category was due to two problems. The first was that of drug names that were not properly classified as entire noun phrases (e.g., "Sudafed Non-Drying Sinus Liquid Caps"). The second was that the phrases 'drug' and 'medication' were not categorised as drugs by the algorithm due to the incorrect dropping of a concept for the ambiguous string 'medicine'. Once this error was resolved, the recall for Drugs and Chemicals increased to 76%, which is 11% higher, and the F-score increased to 72%. These will be added to our custom mappings list to ensure their proper categorisation in the future.

A major source in the 50% recall of the Medical Procedures category was a document on acupuncture, which accounted for one-third of all missed phrases that should have categorised as Medical Procedure. This document contained several unconventional health terms, such as 'adjustment' and 'complementary medicine', that were missed by the algorithm.

7 Conclusion

Our HTO algorithm identifies and categorises health phrases with good precision. This provides an automatic way to create user-friendly content-based, hierarchical categorisation of the information in consumer health web pages. Its dynamic display will be easy to tune to individual user preferences or display types, such as handheld devices with small screens.

Our system does not rely on HTML tags or any manual markup, just on the words and punctuation already present in English text. This, combined with the system's equal performance on web pages from the different consumer health web sites we evaluated, makes it a generic approach that can be applied to many different authors, sites, and document formats. There are many different sources of consumer health information on the internet and we did not want our algorithm to be limited to one in particular.

We will continue to improve the effectiveness of our HTO algorithm, including adding additional stop words and removing incorrect mappings between consumer phrases and clinical phrases. A benefit of using existing systems is that this algorithm will continue to improve as work on its underlying sources (GATE, Consumer Health Vocabulary, SPECIALIST, UMLS) continues. The more accurate each of its components becomes, the more the application will appropriately match the needs of consumers.

Testing the efficacy of the visualisation applet has already begun with seniors, whose memory and visual acuity can be affected by their age, and Hispanics who speak English as a second language and may have difficulty comprehending written complex English texts. We also plan on conduct user testing with hospital in-patients, whose high stress levels can affect their ability to retain information. We are continuing user testing iteratively throughout the application's development, as prior pilot studies have provided substantial feedback that has been incorporated into the current system.

Acknowledgements

This work is supported by the NLM grant R21-LM008860-01, Visualisation of Consumer Health Information. We thank Elizabeth Wood for her early feedback on this system.

References

Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, AMA (1999) 'Health literacy: report of the council on scientific affairs', *Journal of the American Medical Association*, Vol. 281, pp.552–557.

Baker, L., Wagner, T., Singer, S. and Bundorf, M.K. (2003) 'Use of the internet and e-mail for health care information', *Journal of the American Medical Association*, Vol. 289, pp.2400–2406.

Bashyam, V. and Taira, R.K. (2005) *Indexing Anatomical Phrases in Neuro-Radiology Reports to the UMLS 2005AA*, American Medical Informatics Association (AMIA) 2005, Washington DC.

Bennett, D., Rothschild, R. and Schillinger, D. (2003) *Low Literacy High Risk: The Hidden Challenge Facing Health Care in California*, California Health Literacy Initiative, Emeryville, CA.

Berland, G.K. and Al, E. (2001) 'Health information on the internet: accessibility, quality, and readability in English and Spanish', *Journal of the American Medical Association*, Vol. 285, pp.2612–2621.

ConsumerHealthVocab.org (2007) *Open-Access and Collaborative Consumer Health Vocabulary (OAC CHV)*, in Zeng, Q. (Ed.), http://consumerhealthvocab.org

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002) 'GATE: a framework and graphical development environment for robust NLP tools and applications', *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA, pp.168–175.

Doak, C.C., Doak, L.G. and Root, J.H. (1996) *Teaching Patients with Low Literacy Skills*, J.B. Lippincott Company, Philadelphia.

Elhadad, N., Kan, M-Y., Klavans, J.L. and Mckeown, K.R. (2005) 'Customization in a unified framework for summarizing medical literature', *Journal of Artificial Intelligence in Medicine*, Vol. 32, pp.179–198.

Fox, S. and Fallows, D. (2003) *Internet Health Resources*, Pew Internet & American Life Project, Washington, DC.

Fox, S. and Rainie, L. (2000) *The Online Health Care Revolution*, Pew Internet Organization, Washington, DC.

Garbers, S. and Chiasson, M.A. (2004) 'Inadequate functional health literacy in spanish as a barrier to cervical cancer screening among immigrant Latinas in New York City', *Preventing Chronic Disease*, Vol. 1, pp.1–10.

Hazlehurst, B., Mullooly, J., Naleway, A. and Crane, B. (2005) 'Detecting possible vaccination reactions in clinical notes', *American Medical Informatics Association (AMIA)'05*, Washington DC, USA, pp.306–310.

Hepple, M. (2000) 'Independence and commitment: assumptions for rapid training and execution of rule-based POS taggers', *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong, pp.278–285.

Hirst, G., Dimarco, C., Hovy, E. and Parsons, K. (1997) 'Authoring and generating health-education documents that are tailored to the needs of the individual patient', in Jameson, A., Paris, C. and Tasso, C. (Eds.): *User Modeling: Proceedings of the Sixth International Conference (UM97)*, Chia Laguna, Sardinia, Italy, Springer, pp.107–118.

Institute of Medicine (2004) *Health Literacy: A Prescription to End Confusion*, National Academy Press, Washington DC.

Leavitt, M.O. and Shneiderman, B. (2006) *Research-Based Web Design & Usability Guidelines*, in Services, U.S.D.O.H.A.H. (Ed.), US Government Printing Office, Washington, DC.

Leroy, G., Eryilmaz, E. and Laroya, B.T. (2006) *Health Information Text Characteristics*', American Medical Informatics Association (AMIA) 2006, Washington, DC.

Literacyworks (2007) *California Health Literacy Initiative*, http://cahealthliteracy.org/

Mccray, A.T., Burgun, A. and Bodenreider, O. (2001) 'Aggregating UMLS semantic types for reducing conceptual complexity', in Al, V.P.E. (Ed.): *MEDINFO 2001*, IOS Press, Amsterdam, pp.216–210.

Mccray, A.T., Loane, R.F., Browne, A.C. and Bangalore, A.K. (1999) *Terminology Issues in User Access to Web-based Medical Information*, AMIA, Washington, DC.

Miller, T., Leroy, G. and Wood, E. (2006) 'Dynamic generation of a table of contents with consumer-friendly labels', *American Medical Informatics Association (AMIA) Annual Symposium*, Washington DC, pp.559–563.

National Institutes of Health (2003) *The Plain Language Initiative*, http://execsec.od.nih.gov/plainlang/

National Library of Medicine (2007) *Unified Medical Language System*, National Library of Medicine, Bethesda, MD.

Reiter, E., Robertson, R. and Osman, L. (2003) 'Lessons from a failure: generating tailored smoking cessation letters', *Artificial Intelligence*, Vol. 144, pp.41–58.

Riloff, E. and Jones, R. (1999) 'Learning dictionaries for information extraction by multi-level bootstrapping', *Sixteenth National Conference on Artificial Intelligence (AAAI-98)*, Orlando, FL, pp.474–479.

Root, J.H. and Stableford, S. (1999) 'Easy-to-read consumer communications: a missing link in medicaid managed care', *Journal of Health Politics, Policy and Law*, Vol. 24, pp.1–26.

Rudd, R.E., Moeykens, B.A. and Colton, T.C. (1999) 'Health and literacy: a review of medical and public health literature', in Comings, J., Garners, B. and Smith, C. (Eds.): *Annual Review of Adult Learning and Literacy*, Jossey-Bass, New York, Chapter 5, available from http://www.hsph.harvard.edu/healthliteracy/litreview_final.pdf

Sheffield Natural Language Processing Group (2005) *General Architecture for Text Engineering*, 3.0 ed., Sheffield, UK, http://gate.ac.uk/

The Lexical Systems Group (2007) *The SPECIALIST NLP Tools*, National Library of Medicine, Bethesda, MD.

Tse, T. and Soergel, D. (2003) *Exploring Medical Expressions Used by Consumers and the Media: An Emerging View of Consumer Health Vocabularies*, American Medical Informatics Association (AMIA) 2003, Washington, DC, pp.174–183.

van Servellen, G., Brown, J.S., Lombardi, E. and Herrera, G. (2003) 'Health literacy in low-income Latino men and women receiving antiretroviral therapy in community-based treatment centers', *AIDS Patient Care and STDs*, Vol. 17, pp.283–298.

White, S. and Dillow, S. (2005) *Key Concepts and Features of the 2003 National Assessment of Adult Literacy (NCES 2006-471)*, US Department of Education, National Center for Education Statistics, Washington DC.

Williams, M.V., Baker, D.W., Honig, E.G., Lee, T.M. and Nowlan, A. (1998a) 'Inadequate literacy is a barrier to asthma knowledge and self-care', *Chest*, Vol. 114, pp.1008–1015.

Williams, M.V., Baker, D.W., Parker, R.M. and Nurss, J.R. (1998b) 'Relationship of functional health literacy to patients' knowledge of their chronic disease', *Archives of Internal Medicine*, Vol. 158, pp.166–172.

Zeng, Q., Kogan, S., Ash, N. and Greenes, R.A. (2001) 'Patient and clinician vocabulary: How different are they?', in Al, V.P.E. (Ed.): *MEDINFO 2001*, IOS Press, Amsterdam, pp.399–403.

Zeng, Q.T., Tse, T., Crowell, J., Divita, G., Roth, L. and Browne, A.C. (2005) *Identifying Consumer-Friendly Display (CFD) Names for Health Concepts*, DSG-TR-2005-003.