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

Linguistic Analysis: Terms and Phrases Used by Patients in E-mail Messages to Nurses

Yichuan Hsieh[a],Gudrun Audur Hardardottir[b], Patricia Flatley Brennan[c]

[a] *School of Nursing, University of Wisconsin, Madison, Wisconsin, USA*
[b] *College of Nursing, University of Iowa, Iowa City, Iowa, USA*
[c] *Moehlman Bascom Professor, School of Nursing and College of Engineering, University of Wisconsin, Madison, Wisconsin, USA*

Abstract

While some researchers are focusing on mapping free-text within health care fields into controlled vocabularies and classifications, many researchers are focusing on consumers' vocabularies. Using natural language processing (NLP) tools, such as MetaMap, to extract and map into terms in a controlled vocabulary is one way of understanding the pattern of terms used by lay people. Before an NLP tool can be effectively and efficiently used to extract concepts and create machine-understandable interpretations of the data, the appropriateness of the tool needs to be determined. This study aims were to determine the appropriateness of linguistic meaning captured for terms and phrases used by patients in electronic mail messages to nurses, using nursing-specific MetaMap output. Twenty messages were randomly selected from the 241 messages data set. Results indicated that four out of six nursing classification systems captured more than 50 % of the parsed word's linguistic meaning. This study demonstrates that it is possible to automatically extract and capture the linguistic meaning of the terms patient use in their electronic mail messages.

Keywords:

natural language processing, nursing classification systems, nursing

Introduction

The Internet enables lay people a quick access to knowledge about diseases, disease management, health promotion, and wellness. The vast amount of information provided by the Internet, frequently inadequately organized and of questionable quality, often leads to confusion and anxiety for consumers as well as being very time consuming searching for relevant information [1]. Moreover, research shows that the quality of health outcomes is affected if consumer information needs are not met or answered with low quality, inaccurate, and misleading information [2]. Furthermore, much of the quality information that is provided on the Internet is written in such a format that consumers may have difficulty finding and understanding the information they search for. Many studies have found that there are considerable mismatches between the vocabulary consumers use and the health information terminology that consumers search for [3]. Health information retrieval using consumers' natural language usually leads to very poor results because of mismatches due to misspellings, partial words, and abbreviations, or because of semantic issues (e.g. synonyms) [3-8]. One way to empower consumers to make informed decisions about their health care is to increase access to quality health care information online, information that is written in a language understandable to consumers. Health consumers often use chat groups and electronic mail to seek advice and counsel about their health problems. Approaches used to identify whether terms used by health consumers are present in existing health care vocabularies may provide a basis for linking consumers to accurate and appropriate knowledge resources which in turn could improve their understanding of the health concerns they might have. Such approaches can be applied by using natural language processing (NLP) tools, which link bibliographic databases using controlled vocabularies. NLPs are designed to automatically extract coded data from free text and create machine-understandable interpretations of that data. Research on NLP is focused on extracting specific concepts and capturing the semantic meaning from the free-text and mapping it directly into terms in a controlled vocabulary, such as the Unified Medical Language System® (UMLS®). The National Library of Medicine (NLM) has developed Semantic Knowledge Representation tools to assist users in accessing and manipulating the UMLS and its Metathesaurus to make information more accessible [4].

The UMLS is a long-term interdisciplinary research project developed and maintained by the NLM at the National Institutes of Health (NIH). The goal is to facilitate integration and information retrieval from multiple biomedical information sources. The UMLS includes an organized collection and linkage of over 100 clinical and biomedical vocabularies and classifications [9], with 875,255 concepts and approximately 1,815,280 terms[10]. There are three UMLS Knowledge Sources one of which, the Metathesaurus contains and interconnects multiple clinical and biomedical vocabularies. Currently, six of the nursing classification systems recognized by the American Nurses Association (ANA) are included in the UMLS Metathesaurus:

1.North American Nursing Diagnosis Assoc., 1999 (NAN99)
2.Nursing Interventions Classification, 1999 (NIN99)
3.Nursing Outcomes Classification, 1997 (NOC97)
4.Omaha System, 1994 (OMS94)
5.Home Health Care Classifications, 1996 (HHC96)
6.Patient Care Data Set, 1997 (PCD97)

The Semantic Network is a network of general categories to which all concepts in the Metathesaurus have been assigned.

The SPECIALIST lexicon is intended to provide lexical information needed for the SPECIALIST Natural Language Processing System and includes both biomedical terms as well as commonly occurring English words. The lexical tools address the high quantity of inconsistency that occurs in natural language words and terms. Several lexical programs are available with the UMLS Knowledge Sources for searching, indexing, and lexical processing [10]. One such tool is the MetaMap program, which was developed to map free text to a biomedical knowledge source to improve information retrieval and recently also to identify Metathesaurus concepts referred to in texts [11].

According to Aronson [11], there are five steps involved in the MetaMap processing. First, text is parsed into simple noun phrases using the SPECIALIST minimal commitment parser. Variants are then generated for each phrase using knowledge in the SPECIALIST lexicon and a supplementary synonyms database. A variant includes a term as well as all its synonyms, acronyms, abbreviations, derivational variants, any meaningful combinations of these, and spelling variants. All Metathesaurus strings or candidates that contain at least one of the variants are retrieved, and then evaluated against the input text. Evaluation consists of a) computing a mapping from the phrase words to the candidate's words, and b) calculating the strength of the mapping using weighted average of four metrics, centrality, variation, coverage, and cohesiveness [11,12]. For each of these four components, a normalized value between 0 (the weakest match) and 1 (the strongest match) is calculated. The MetaMap candidates are then ordered in relation to the strength of the mapping. Finally, the strength of the complete mapping is calculated by computing a weighted average of the four components to a normalized value between 0 (indicating no match at all) and 1000 (indicating a perfect match), and the highest scoring complete mappings are chosen as MetaMap's best interpretation of the original phase [13]. However, it is very unlikely that good mapping results will be accomplished when a complex Metathesaurus string is parsed as MetaMap processing involves the parsing of text into simple noun phrases [14].

The purpose of this study was to determine the appropriateness of linguistic meaning captured for terms and phrases used by lay people in e-mail messages to nurses, using nursing-specific MetaMap output. The current study focused on two questions: a) How well did the suggested MetaMap term capture the linguistic meaning of the parsed term? b) Which nursing vocabulary captured the linguistic meaning of the parsed term most often?

Materials and Methods

Data source

The data set consisted of 241 electronic messages sent from patients to the nurse in the "HeartCare project", a project that provided consumers recovering from Coronary Artery Bypass Graft (CABG) surgery with communication applications and support network through a standard Web browse [4]. All unique identifiers had been extracted from the messages. The MetaMap output was in Excel format as well as text file format using Standardized Nursing Languages as source text.

Method

Every tenth parsed message of the 241 messages was randomly retrieved and analyzed, beginning with message #10 and ending with message # 200, for a total of 20 messages. However, if the selected tenth message was a nurse-written message, the next patient-written message was selected instead of using the nurse-written message to maintain message selection criteria. The selected messages were copied into a Word file. All parsed phrases from these messages (aggressive run which covers extraneous words) were retrieved and put into a separate Word file. Then all recognized meta candidates were copied and put into a table in a third Word file. The data analysis table included 6 columns; one for the parsed term/phrase from the original message, one for the score of the MetaMap candidates, one for the final MetaMap given candidate score, one for mapping terms within the nursing vocabularies, one for mapping nursing vocabulary, and one Yes/No column for appropriateness of term recognition (i.e., how well the linguistic meaning of the parsed term was recognized). Table 1 illustrates what the table would look like for one meta candidate in our data analysis.

All messages were validated "blindly" by researchers independently. The messages were read only after finishing the analysis on the appropriateness of word recognition of the parsed terms/ phrases. To check inter-rater reliability, all messages were checked by both researchers independently. Disagreements were later discussed and resolved. Inter-rater reliability was 98%.

Analysis

The mappings of the parsed terms and phrases to the nursing vocabularies were examined and appropriateness judged, based on the individual judgment of the researcher. No fixed set of rules was used as these cannot generally capture all the possible meanings of language. However, some guidelines were established:

1. Mapping was considered to be appropriate if it pertained to the general meaning and most common usage of a word: i.e., the word "read" was appropriately mapped as "spiritual reading" and "inability to read" but "abnormal blood pressure reading" was considered not to be appropriate as this refers to a different function.
2. The most likely meaning or usage of a term or phrase was taken into account, thus "the information changed as far" was deemed to map accurately to "information disclosure" and "information management", but the mapping term "voice changes" (voice changes due to physical maturation) is clearly not appropriate.
3. All possible MetaMap candidates were included in the analysis to examine whether "lower score" candidates could possibly be more appropriate than the "higher score" candidates that the MetaMap program chooses as most appropriate candidates.

Twenty messages were extracted and analyzed. A total number of 1,305 terms and phrases were analyzed. Descriptive statistics using frequency tables, percentage, and bar graphs were used as applicable.

Table 1: Example of data analysis

Parsed term/phrase from message	Meta mapping candidates score	Meta mapping score	Mapping term within a nursing vocabulary	Mapping nursing vocabulary	Appropriateness Yes/No
smoking	1000	1000	Smoking	OMS94 [Individual Behavior]	Yes
			Smoking Cessation Assistance	NIC99 [Educational Activity, Therapeutic or Preventive Procedure]	Yes
			Smoke detector maintenance	NOC97 [Finding]	No
			Description of use of functioning smoke detectors	NOC97 [Finding]	No

Results

The total number of candidate term/phrase recognized by MetaMap in the 20 messages was 1,305 out of 162 sentences and 2,147 words in the messages. The number of unique mapping terms from each nursing vocabulary is shown in Table 2. For example, there were 55 terms/phrases which matched to NAN99, and 31 matched terms/phrases were determined by the researchers to be adequate. The source vocabulary that produced the highest number of matches was NOC97 with 523 matches. NIC99 came second, followed by HHC96, OMS94, PCDS97, and then NAN99.

For the mapping appropriateness, a total of 692 matches were determined to be appropriate. The mean of the appropriate matching rate was 53.03%. Half of the six nursing classification systems (NOC97, OMS94, and NAN99) captured linguistic meaning above average. A total of 311 appropriated matches were found in the NOC99, which yielded the best matching rate of 59.46%. Results also indicated that four out of six nursing classification systems (NOC97, OMS94, NAN99, and PCDS97) captured more than 50 % of the parsed term/phrase's linguistic meaning. The detailed summary of results for all nursing classification systems are presented in Table 2 and Figure 1.

Table 2. Mapping Results by Nursing Vocabularies

	Total number of term/ phrase matches	Total number of appropriate term/ phrase matches	Total % of appropriate term/ phrase matches
NAN99	55	31	56.36
NIC99	320	156	48.75
NOC97	523	311	59.46
OMS94	132	77	58.33
HHC96	179	70	39.11
PCDS97	93	47	50.54
Total	1305	692	53.03

Discussion

The NOC97 appears to have both the most total numbers of terms/phrases matched as well as the highest rate of appropriate matches. The NOC is a classification vocabulary on nursing sensitive patients' outcomes and, as such, should be likely to capture terms that describe patient's recoveries or outcomes, better. In this project, the NOC did show some promise in capturing linguistic meanings of terms/phrases used by patients in free text messages.

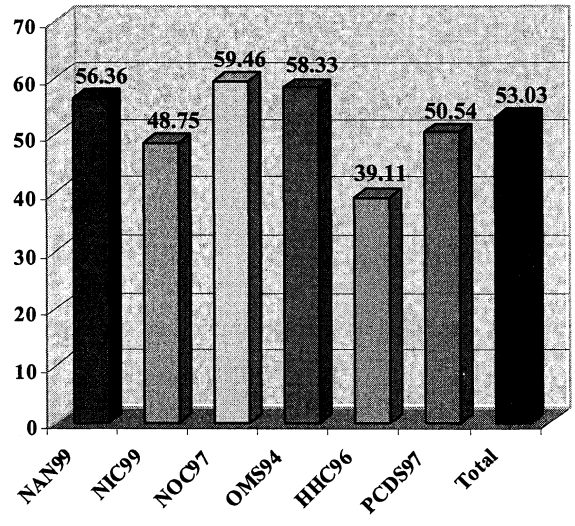

Figure 1 - Percentage of appropriate term/phrase matches

The Omaha System (OMS) and the Home Health Care classification system (HHC) were both developed with a public health care focus and include both interventions and patient assessment concepts. Hence, it would be expected that they would capture these patients' terms. Therefore, it is surprising that the HHC only captured 39.11% of the parsed terms' linguistic meanings. NIC is a nursing vocabulary developed to capture nursing inter-

ventions, NANDA was developed to capture patient problems, and hence, how well they matched is not unexpected. On the other hand, the Patient Care Data Set (PCDS) is derived from inpatient care. Therefore, one should anticipate lower capturing rate in the data set used for this study; however, there doesn't seem to be a significant difference between the appropriateness of term/phrase matches by the three nursing vocabularies, NOC, NANDA, and the Omaha System, all have close to 60% of mapping accuracy.

In this study, a match was considered to be proper (true positive) whenever the MetaMap detected a source document term/phrase and mapped it appropriately to an UMLS term. For example, the parsed phrase "my pain" was mapped to "pain", "chronic pain" and "acute pain". This was considered a proper match.

One of the most common problems with the MetaMap program is ambiguity because the program cannot always distinguish between words[11]. This study found three types of error with the MetaMap program:

1. "False positives" in which the terms and phrases mapped incorrectly to an UMLS term. In this study the word "smoking" mapped appropriately to "smoking", the phrase "smoking cessation assistance" was also considered an appropriate match, although it is an "overmatch" as it includes words at the end of the "string" that do not participate in the match. However, the words "smoke detector maintenance" and "description of use of functioning smoke detectors" (also an overmatch), were considered as "false positive", as the terms mapped incorrectly to a nursing vocabulary.
2. "False negatives" in which terms and phrases that appear in the message are not recognized by the parser. Several terms were considered to be "false negatives." Those included such terms as "dizziness", "dizzy", "cardiac rehab" and "shortness of breath," to name a few.
3. "Vocabulary insufficiency" in which terms/phrases are correctly recognized by the parser but lack a concept within the vocabulary classification. For example the phrase "at home" was mapped inappropriately to "homelessness", "impaired home maintenance management", "home maintenance assistance", and "home situation analysis".

Conclusion

In order to empower consumers to make informed decisions regarding healthcare, it is essential that available health information be written in a language easily understandable to the most people. Our results show that the MetaMap program (using nursing specific classification systems) captured the linguistic meaning of the parsed terms used by the patients in this project 53.03% of the time. This finding demonstrates that it is possible to use existing NLP tools to automatically extract and capture the linguistic meaning of the terms patients use in their electronic mail messages, which is illustrated by identifying terms that are found in standard health care vocabularies. This information can be used to improve patients' access to quality health care information on the Internet and other electronic resources.

While our findings suggest the NOC performed the best among all six nursing classification systems, one should be mindful to make definite conclusion about which nursing classification system is the best one to describe or capture terms/phrases used by patients, as the nursing vocabularies were developed for different purposes. Moreover, different classification systems together could enhance the precision of the mapping capacity.

Again, this study's aims were to explore the level of the appropriateness of an existing NLP tool, MetaMap, solely in capturing linguistic meaning of the terms used by patients in free text messages, without considering the term/phrase meaning within the context of a message. Therefore, this study is just the starting point. Further work examining the relationship within the context of a message would advance knowledge in the NLP development.

Acknowledgements

We would like to thank Dr. Alan R. Aronson for his help on MetaMap data preparation. Also we would like to thank the UW-Madison Health Systems Lab for its support and comments on this paper. Yichuan Hsieh is supported in part by the NIH/NINR T32 grant (T32NR07102).

References

[1] Eysenbach G, Jadad AR. Evidence-based patient choice and consumer health informatics in the Internet age. Journal of Medical Internet Research 2001; 3(2): e19.

[2] Boulos MNK, Roudsari AV, Carson ER. A dynamic problem to knowledge linking Semantic Web service based on clinical codes. Medical Informatics & the Internet in Medicine 2002; 27(3): 127-137.

[3] Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods of Information in Medicine 2002; 4(41): 289-298.

[4] Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in E-mail. Journal of Biomedical Informatics (in press).

[5] Kogan S, Zeng Q, Ash N, Greenes RA. Problems and challenges in patient information retrieval: a descriptive study. Proceedings / AMIA Annual Fall Symposium. 2001; 329-33.

[6] McCray AT, Loane RF, Browne AC, Bangalore AK. Terminology issues in user access to Web-based medical information. Proceedings/AMIA Annual Fall Symposium. 1999; 107-11.

[7] Patrick TB, Monga HK, Sievert ME, Houston Hall J, Longo DR. Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. Journal of Medical Internet Research 2001; 3(3):E24.

[8] Zeng Q, Kogan S, Ash N, Greenes RA. Patient and clinician vocabulary: how different are they? MedInfo 2001; 10(Pt 1):399-403.

[9] NLM (2003) Metathesaurus B3. Retrieved March 20, 2003 from: http://www.nlm.nih.gov/research/umls/METAB3.HTML

[10]NLM (2003). Fact Sheet. UMLSR Metathesaurus. Retrieved March 20, 2003 from: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

[11]Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proceedings / AMIA Annual Fall Symposium. 2001; 17-21.

[12]Aronson AR, Boedenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM Indexing Initiative. Proceedings / AMIA Annual Fall Symposium. 2000; 17-21.

[13]Aronson, A.R. (2001). MetaMap Evaluation. Retrieved March 30, 2003 from: http://skr.nlm.nih.gov/papers/MetaMap

[14]Aronson, A.R. (2003). Filtering the UMLS® Metathesaurus® for MetaMap (2002 Edition). Retrieved March 30, 2003 from: http://skr.nlm.nih.gov/papers/MetaMap

Address for correspondence

Yichuan Hsieh, MS, RN
H6/296 Clinical Science Center
School of Nursing, 600 Highland Avenue
University of Wisconsin
Madison, WI 53792, USA.
E-mail: yichuanhsieh@wisc.edu