Evaluating UMLS Strings for Natural Language Processing

Alexa T. McCray[a], Olivier Bodenreider[a], James D. Malley[b], Allen C. Browne[a]

[a]National Library of Medicine
[b]Center for Information Technology
National Institutes of Health

The National Library of Medicine's Unified Medical Language System (UMLS) is a rich source of knowledge in the biomedical domain. The UMLS is used for research and development in a range of different applications, including natural language processing (NLP). In this paper we investigate the nature of the strings found in the UMLS Metathesaurus and evaluate them for their usefulness in NLP. We begin by identifying a number of properties that might allow us to predict the likelihood of a given string being found or not found in a corpus. We use a statistical model to test these predictors against our corpus, which is derived from the MEDLINE database. For one set of properties the model correctly predicted 77% of the strings that do not belong to the corpus, and 85% of the strings that do belong to the corpus. For another set of properties the model correctly predicted 96% of the strings that do not belong to the corpus and 29% of the strings that do belong to the corpus.

INTRODUCTION

The 12th edition (2001) of the Unified Medical Language System® (UMLS®) Metathesaurus collects terms from over 50 biomedical vocabularies [1]. Each of these vocabularies was created and is maintained for a variety of purposes, including use in patient record systems, in billing systems, and for indexing the biomedical literature. It is not surprising, then, that not all Metathesaurus strings are suitable for natural language processing (NLP) applications. The objective of this study is to define and evaluate methods whereby individual UMLS strings can be selected for their usefulness in NLP applications.

Medical language processing is an active area of research, and recent developments hold some promise, particularly in specific application areas [2,3]. All NLP systems need access to robust lexical knowledge, which is not always readily available, although resources such as the UMLS offer some help [4,5]. If the terms that are used in a natural language corpus are found in the UMLS, then the NLP system has access to extensive domain knowledge as well [6-8].

METHODS

We used the occurrence of a string in a natural language corpus as an indicator that it will prove useful for natural language processing. We drew our corpus from the National Library of Medicine's (NLM's) MEDLINE® bibliographic database. MEDLINE includes citations to articles in over 4,000 journals, broadly covering biomedical research and the clinical sciences, including nursing, dentistry, veterinary medicine, pharmacy, allied health, and pre-clinical sciences. We used a corpus that represents all the citations entered into MEDLINE during 1999. We used the titles and abstracts in this corpus of 439,741 citations; 78% of the citations included abstracts.

The 2001 release of the Metathesaurus has 1,457,129 English strings, organized into 797,359 concepts. We merged strings that differed only by case, giving us a total of 1,397,429 unique strings. We matched each of these strings against the corpus, retaining all string features, (e.g., punctuation, spacing, word order) with the exception of case.

Further, we identified several properties that we hypothesized would serve to classify strings in the Metathesaurus as either useful or not for NLP. Using these properties, we would then be able to predict the likelihood of a given string being found in a target corpus, as well as to predict the strings that are not likely to be found in the corpus. The overall goal is to develop a set of predictors that would allow us to filter out ill-formed strings for NLP applications. We selected a total of fifteen properties for our experiment. These are shown in Table 1 in the Appendix and include a description, some examples, and the number of strings in the Metathesaurus that have that property.

1067—5027/01/$5.00

The majority of the properties we identified relate in some way to the form of the string and are likely not to be found in natural written or spoken English. For example, permuted terms, found in some controlled vocabularies for browsing and look-up purposes (e.g., "blood pressure, abnormal") do not reflect the way medical concepts are expressed in natural language corpora. We included a property called CT_COMMA_SP (contains comma followed by a space) to mark these cases. For terms that include phrases such as "not elsewhere classified", "NEC", or "without mention of" we included a property called ANY_CLS (any classification feature). In order to identify the Metathesaurus strings that have the properties we identified, we used regular expressions. For example, the regular expression for the property CT_NUM (contains a number) is '/[0-9]/'.

All properties are binary with the exception of NB_SOURCES and NB_WORDS. NB_SOURCES counts the number of sources in which the string appears. The UMLS documentation [1:132-70] lists some one hundred source abbreviations, naming the vocabularies included within the Metathesaurus. In some cases, there are several historical versions of the same vocabulary. For example, there are four versions of the COSTAR vocabulary, representing releases in 1989, 1992, 1993, and 1995. For the purposes of this work, we consider these a single source 'family' and count it as one source. There are 56 source families in the 2001 Metathesaurus. The sources vary significantly in their scope, structure, and in the nature of the strings they contain. They include terminologies that cover specific areas such as substance abuse, adverse reactions, and nursing to more broadly based terminologies, including those used for billing purposes. The number of strings in a vocabulary varies from as small as 43 for the Glossary of Methodologic Terms for Clinical Epidemiologic Studies of Human Disorders to as large as 467,535 for the Medical Subject Headings.

NB_WORDS counts the number of words in a string. We compared several sources, including the SPECIALIST lexicon, Dorland's Illustrated Medical Dictionary [9] Webster's Dictionary [10], and the UMLS Metathesaurus for the distribution of words in a term. It is likely that a large percentage of Metathesaurus strings will have more words than those found in standard dictionaries, and, therefore, may also not be found in free text.

The remaining properties are derived from the term type labels that have been applied to strings as part of the process of building the Metathesaurus. These labels are source specific and are attributes of the particular name in that vocabulary. The term type is found in the TTY field of the MRSO file, and each type is defined in the UMLS documentation [1:141-3]. We studied the set of 96 term types, identified those that we thought might be useful for our purposes, and then grouped them according to shared characteristics. As an example, TTY_SHORT_FORM groups nine term types that indicate that the string is a shortened form, such as an abbreviation or truncated form. TTY_PHRASE groups several term types that are used for nursing activities. The strings that are marked in this way in the Metathesaurus are more akin to instructions than they are to terms that might be found in a natural language corpus.

Since it seemed unlikely that a single property would be sufficient as a predictor of which strings would be useful for NLP, and which would not, and since there is no obvious combination of predictors based on *a priori* knowledge, we used statistical techniques to help us determine a combination of predictors that would achieve our goal.

From a statistical perspective, this task can be formulated as a classification problem, in which the predictors are used to determine the value of a binary target variable. The method of choice for achieving such a classification task with good estimates of the misclassification error rates is a nonparametric, tree-structured approach called binary recursive partitioning with cross-validation [2]. We should note that standard estimates of these rates using observed misclassifications or even the popular leave-one-out approach are known to be consistently biased in an optimistic direction. Using 10-fold cross-validation instead provides considerably more accurate error rate estimates.

For example, when used to generate a classification scheme, and given a set of predictors, A and B, and a target variable (appearance in the MEDLINE corpus, in our problem), generation of a binary tree begins by considering all splits of the data into two pieces based on the possible values of A, the first predictor. Let us consider that predictor A represents the fact that a string contains a digit. A has two states, marked as 'yes' if the string contains a digit and marked as 'no' if it does not. Similarly, the target variable has two states, marked as 'yes' if the string appears in the target corpus and marked as 'no' if it does not.

The best splitting rule using A is determined by minimizing the within group sums of squares, when one state is assigned the numerical value 0, and the other is assigned the value 1. A similar optimal split

is found using property B. The sum of squares obtained using the splitting rule based on B is compared with the sum of squares obtained by splitting on A. The optimum first split (the choice of the predictor and the splitting value for that predictor) is then found. Each partitioning of the sample space is then repeatedly considered for additional partitions, by selection over the predictors and choices of splits.

We used the CART software package [12] to carry out the statistical analysis. For technical reasons imposed by the software, we created three randomized subsets of the full set of Metathesaurus strings, two sets of 470,000, and the third set of 457,429. We ran the experiment separately on each of these sets, using all 15 variables and checked for convergence in the results.

RESULTS

Mapping the entire set of Metathesaurus strings to the corpus, resulted in a 10% match. A total of 144, 396 of the 1,397,429 strings were actually found in MEDLINE. This means that fully 90% of the strings were not found. There were a few cases in which a string matched MEDLINE text and was incorrectly counted as a match. For example, we were surprised to see that a string like "depression, psychotic" was found in MEDLINE. On further investigation this turned out to be a false hit, having matched the text "Fifty-three percent of the total sample were found to be affected by one or more psychopathological problems; the most frequent were *depression, psychotic* disorders, cognitive disturbances ..." The mapping method involved a simple string match and, as a result, these cases introduced a small amount of noise in the sample.

The average number of words in a string found in the lexicon and in Webster's is one word. The average found in the corpus and in Dorland's dictionary is two words, and the average for the Metathesaurus is five words. Perhaps more interesting is to compare the percentage of strings that have more than, for example, three words in each of these sources. For Webster's this is essentially zero (.003%), for the lexicon it is 2%, for the strings found in the corpus it is 8%, for Dorland's it is 13%, and for the Metathesaurus it is more than half (53%).

We were able to get excellent convergence among the three randomized subsets of Metathesaurus strings when running the 15 properties against the target variable. The top four properties were common to all three subsets and the percentage of well-classified strings was similar for all subsets.

See Figure 1 for an illustration of the tree that CART builds as it generates the classification scheme.

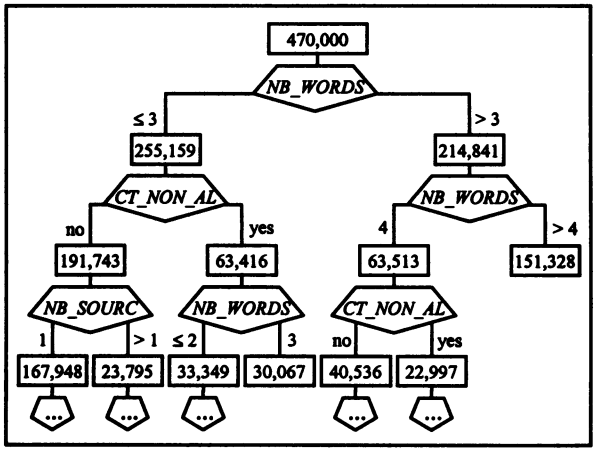

Figure 1 – Top nodes of the classification tree for one subset of 470,000 strings (rectangles contain the number of strings after each split; pentagons contain the name of the variable used for splitting)

The classification process can be summarized by the following two indicators:

- *Sensitivity*, which represents the percentage of well-classified strings that do not belong to the corpus (or the probability of predicting that a string does not belong to the corpus, given that it does, in fact, not belong to the corpus), and

- *Specificity*, which represents the percentage of well-classified strings that do belong to the corpus (or the probability of predicting that a string belongs to the corpus, given that it does, in fact, belong to the corpus).

When we put all 15 properties in the model, the model correctly predicted :

- 77% of the strings that do not belong to the corpus
- 85% of the strings that do belong to the corpus

The top four predictive properties were, in order, NB_WORDS, CT_NON_ALPHN, ANY_PAREN, TTY_SHORT_FORM. Based on these results we decided to process just these four properties in the CART system. In this case, the model correctly predicted:

- 67% of the strings that do not belong to the corpus

- 91% of the strings that do belong to the corpus

The property NB_WORDS when used alone made similar predictions (69% and 81%, respectively).

Since our primary goal is to develop methods for filtering the Metathesaurus, we experimented with small sets of properties to see if we could improve our predictions for the strings that do not belong to the corpus. The four properties NB_SOURCES, CT_AND_OR, ANY_UNSP, and ANY_CLS correctly predicted:

- 96% of the strings that do not belong to the corpus
- 29% of the strings that do belong to the corpus

Interestingly, the property NB_SOURCES when used alone predicted equally as well.

DISCUSSION

The results reported here are indicative, rather than conclusive. The properties we have chosen to investigate hold some promise for identifying those strings that are likely to appear in natural language text. Our preliminary results should, however, be carefully interpreted. First, we looked at only one corpus, representing only one year of the MEDLINE database. Although the corpus is large, there are still some legitimate words that did not happen to appear during that year (e.g., "saltpeter", "xerography"). Second, there are undoubtedly other string properties that may be of interest and that may have an impact on the overall results.

For the corpus and properties we did use, we are able to draw some preliminary conclusions. Both the number of words in a string, and the number of sources in which a string appears, are important predictors of the "goodness" of a string for NLP purposes. The longer the string is, the less likely it is to be found in a corpus, and, therefore, the less likely it is to be useful for natural language processing, and if a string appears in several sources, then it is more likely to reflect a standard way of expressing a concept and therefore more likely to be useful for language processing. The term types did not have as much predictive power when used with the other properties, but further experimentation is needed.

The methodology and results described here are the first steps in our longer-term effort to develop methods to filter the large and complex Metathesaurus for natural language processing purposes. The UMLS is a rich source of knowledge for the biomedical domain. The extent to which NLP applications are able to take advantage of that knowledge depends in part on the extent to which they are able to map natural language text into the UMLS construct. An estimate of the percentage of strings in a particular source that do belong to a corpus may also be helpful in evaluating the usefulness of that vocabulary for NLP purposes.

We will continue our experimentation by varying the number and nature of properties considered, using our *a priori* knowledge of the nature of the terminologies included within the UMLS, as well as through further statistical analysis.

REFERENCES

1. National Library of Medicine. Documentation, UMLS Knowledge Sources. 12th edition, January 2001.
2. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med. 1999 Aug;74(8):890-5.
3. Spyns P. Natural language processing in medicine: an overview. Methods Inf Med 1996 Dec;35(4-5):285-301.
4. McCray AT. The nature of lexical knowledge. Methods Inf Med 1998 Nov;37(4-5):353-60.
5. Aronson AR. The effect of textual variation on concept based information retrieval. Proc AMIA Annu Fall Symp 1996;:373-7.
6. Johnson SB. A semantic lexicon for medical language processing. J Am Med Inform Assoc 1999, May-Jun;6(3):205-18.
7. Hahn U, Romacker M, Schulz S. How knowledge drives understanding – matching medical ontologies with the needs of medical language processing. Artif Intell Med. 1999 Jan;15(1):25-51.
8. Baclawski K, Cigna J, Kokar MM, Mager P, Indurkhya B. Knowledge representation and indexing using the Unified Medical Language System. Pac Symp Biocomput. 2000;:493-504.
9. Dorland's Illustrated Medical Dictionary, 27th edition. WB Saunders Company, 1988.
10. Webster's New International Dictionary of the English Language, second edition, C. G. Merriam Company, 1959.
11. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. 1984. Chapman & Hall/CRC Press.
12. Steinberg, D, Colla P. CART system software manual. 1997. Available from Salford Systems (www.salford-systems.com).

APPENDIX

Property	Description	Examples	UMLS Strings
ANY_CLS	Any classification feature, e.g., "other" at the beginning of a string, "not elsewhere classified", "NEC", "without mention"	Unclassified tumor, benign Speech Disorders Not Elsewhere Classified	27944
ANY_UNSP	Any underspecification feature, e.g., NOS, "not specified", "unspecified", "not otherwise specified"	Tic disorder, NOS Bacterial infection, unspecified, in conditions classified elsewhere and of unspecified site	61995
ANY_PAREN	Any bracketed expression, i.e., the string contains an expression enclosed by brackets, parentheses	Dysthymia (or Depressive neurosis) Full-thickness skin loss due to burn [third degree NOS] of foot	148411
CT_COMMA_SP	Contains a comma followed by a space (often an 'inverted' string)	Yellow fever, jungle Sweating, absent	238012
CT_NON_ALPHNM	Contains at least one non-alphanumeric character (dash, dot, apostrophe, space are grouped with alphanumeric)	Oral/nasal mucosal ulcers Weight loss >=10% of body weight	506820
CT_NUM	Contains at least one digit	1, 2-Diacylglycerol Chromosome 5	376112
CT_AND_OR	Contains, but does not start or end with, "and", "or", "and/or"	Larynx and pharynx Hemorrhoidectomy, internal and external, complex or extensive	70573
NB_SOURCES	Number of vocabularies in which the string is found	"Aleutian disease" appears in 2 sources "Heart" appears in 14 sources	1397429
NB_WORDS	Number of words in the string	"Chronic rhinitis" consists of 2 words "Adjustment disorder with mixed disturbance of emotions and conduct" consists of 9 words	1397429
TTY_CHEMICAL	Chemical names (Includes term types N1, NM, CE)	CY 222 Cytidine cyclic 2,3 monophosphate	318078
TTY_LOINC	LOINC complex names (Includes term types CN, CX, LN, LO, LS, LX, SX)	ACIDITY.TITRATABLE ADENINE:MASS:POINT IN TIME:DOSE MED OR SUBSTANCE:QUANTITATIVE	62571
TTY_METADATA	Strings starting with a code or ending with a polysemy marker (Includes term types HX, MM)	A64-A65 AGNOSIAS Blood <2>	18214
TTY_PHRASE	Strings that are generally not noun phrases; often they are full utterances (Includes term types AC, CL, GO, OR, SA, TA)	Patient will adhere to special diet Adjust environment (e.g., light, noise, temperature, mattress, and bed) to promote sleep	11576
TTY_PRESCRIP	Fully specified names for branded drugs, supplies, often including dosage (Includes term types BD, CD, MS)	Tobradex, 0.1%-0.3% ophthalmic ointment Ensure Plus	62201
TTY_SHORT_FORM	Abbreviations, truncated strings (Includes term types AA, AB, CS, DS, ES, NS, OA, PS, SN)	2-malig neop LN head/face/neck HACBP	126399

Table 1: List of Properties