Evaluating the UMLS as a Source of Lexical Knowledge for Medical Language Processing

Carol Friedman[1,2], Hongfang Liu[3], Lyuda Shagina[2], Stephen Johnson[2], George Hripcsak[2]

[1]Computer Science Department, Queens College CUNY
[2]Department of Medical Informatics, Columbia University
[3]Computer Science Department, CUNY Graduate Center

Medical language processing (MLP) systems rely on specialized lexicons in order to recognize, classify, and normalize medical terminology, and the performance of an MLP system is dependent on the coverage and quality of such lexicons. However, the acquisition of lexical knowledge is expensive and time-consuming. The UMLS is a comprehensive resource that can be used to acquire lexical knowledge needed for medical language processing. This paper describes methods that use these resources to automatically create lexical entries and generate two lexicons. The first lexicon was created primarily using the UMLS, whereas the second was created by supplementing the lexicon of an existing MLP system called MedLEE with entries based on the UMLS. We subsequently carried out a study, which is the primary focus of this paper, using MedLEE with each of the two lexicons and also the current MedLEE lexicon to measure performance. Overall accuracy, sensitivity, and specificity using the lexicon primarily based on the UMLS were .86, .60, and .96 respectively. Those measures using the MedLEE lexicon alone were .93, .81, and .93, which was significantly better except for specificity; performance using the supplemental lexicon was exactly the same as performance using solely the MedLEE lexicon.

INTRODUCTION

In order to function properly, medical language processing (MLP) systems rely on specialized lexicons to recognize, classify and normalize clinical terms that are found in textual reports. Manual construction of a lexicon is labor intensive, requires medical expertise, and often forms a bottleneck in the MLP development process. The Unified Medical Language System (UMLS)[1, 2] is a comprehensive source of knowledge in the medical domain that could be useful for acquiring the information needed to automatically generate lexical entries.

Several research groups have explored the acquisition of lexical knowledge for MLP purposes using external knowledge sources[3-5] in order to reduce the effort. In this paper we acquire lexical knowledge using the UMLS. To test the usefulness of the automatically acquired lexicon we evaluated the performance of an existing MLP system, called MedLEE [6], when using the lexical entries that were automatically generated from the UMLS. For this study we utilized a test set of reports and corresponding reference standard from a previous study.

BACKGROUND

Several groups exploited use of controlled medical terminologies for the purpose of facilitating acquisition of lexical knowledge for use in medical language processing. Baud[3] explored acquisition of linguistic lexical knowledge from different language versions of ICD10, a mono-axial system associated with disease terms. Similarly, Zweigenbaum[4] explored the use of SNOMED, a multi-axial semantic system, for acquiring lexical knowledge for MLP in French. Johnson[5] automatically created a semantic lexicon in English using the UMLS. Acquisition methods we describe in this paper are based on his work. The above papers primarily focused on the acquisition methodology and subsequent evaluation of the coverage of the resultant lexicons. Although our paper also describes an automated method for acquiring lexical knowledge from the UMLS, our primary goal involves measuring the performance of an MLP system when actually using the lexicons that were created.

The UMLS Meta incorporates a comprehensive collection of medical concepts and synonymous ways of expressing the concepts; each distinct concept is assigned a unique concept identifier, and all strings, called concept names, that correspond to that concept are assigned the same concept identifier. The UMLS Semantic Network[7] consists of a semantic classification system for the concepts represented in the Meta. Each concept identifier in the UMLS is assigned one or more semantic classes. The Specialist lexicon identifies single- and multi-word phrases along with their variant forms, syntactic classifications, and base forms.

MedLEE[6] is an MLP system that extracts, structures, and encodes relevant clinical information that occurs in patient reports. It is written in Quintus Prolog and

1067—5027/01/$5.00 © 2001 AMIA, Inc.

can run on Windows and most Sun platforms. MedLEE contains several knowledge-based components, one of which is the lexicon. In MedLEE, a lexical entry consists of the word or phrase being defined, the semantic category, and the target form, which is generally the normalized form (e.g. the target for *abdominal* is **abdomen**) The MedLEE system incorporates 55 different semantic classes whereas the 2000 version of the UMLS semantic network has 188 classes. The scope of the UMLS semantic classes is broader than MedLEE's because the UMLS contains categories that are not clinically oriented and therefore not covered in MedLEE. The granularity of the two sets of classes is also different. Accordingly, some of the classes in MedLEE are finer grained than corresponding UMLS classes and vice versa. For example, in MedLEE there is a class only associated with severity information (e.g. *mild*), whereas in the UMLS the class that contains *mild* also contains *beauty* and *developing country.*

The evaluation study described in this paper is based on a prior study [8]; we already had available a training set, a test set and a reference standard from that study. The study consisted of an evaluation of an automated system that computed the risk classification of patients with community-acquired pneumonia by relying on a score that was computed using variables such as certain comorbidities (e.g. liver failure), vital signs, laboratory tests, and demographic information. Some of the variables (i.e. certain laboratory values and demographic information) were available in coded form from the clinical repository at New York Presbyterian Hospital. However, most of the variables were determined by first processing narrative discharge summaries and chest x-ray reports using MedLEE to obtain structured output. The final values for these variables were then obtained by executing queries based on the structured output generated by MedLEE. For example, one way for a query to determine the presence of fever consisted of looking for a finding **temperature** whose value was 101 or more. An expert wrote the queries after analyzing target output values produced by MedLEE subsequent to the processing of the training set. A reference standard was established for the text-based variables of the test set using another independent medical expert who manually read the reports and determined the values. Performance measures for the automated system were computed by comparing the results obtained by the system to those generated by the reference standard.

METHODS

This section consists of two components. The first component describes the method used to automatically generate the two lexicons for the study. One lexicon, called LUMLS, contained clinical entries based solely on the UMLS; the second lexicon, called M+UMLS, contained the MedLEE lexicon supplemented by entries based on the UMLS. The MedLEE lexicon used for the previous study[8] is represented in this paper as M-PRV. Because, MedLEE has been refined and expanded since then, the more recent lexicon, which was used for the current study, was different and is represented here as M-CUR. The second component of this section describes the evaluation that was performed by processing text reports using each of the different lexicons, identifying the values of variables based on the output generated by MedLEE, and comparing the values to the reference standard.

UMLS List of Concept Names: A UMLS list of concept names was generated and used as a basis for creating LUMLS and M+UMLS. In order to create the list we first identified the UMLS semantic categories that were appropriate to use for generating lexical entries for the MedLEE system without any manual intervention. A separate file was created for each UMLS category containing all concept names that corresponded to that category. The UMLS categories were then analyzed by looking at their definitions and samples of the corresponding concept names. We identified four types of UMLS categories: i) those that corresponded exactly to a category in MedLEE, such as **pharmacologic substance,** ii) those that corresponded to a broader category in MedLEE (i.e. **anatomic abnormality** corresponded to **finding** in MedLEE), iii) those that had no correspondence in MedLEE, such as **research activity,** and iv) those that contained a majority of concepts that could not be reliably mapped to a single MedLEE category. For example, **laboratory or test result** contained UMLS concept names that could be mapped in MedLEE to **labtest** (e.g. *prothrombin time*), **bodymeas** (e.g. *wedge pressure*), or **finding** (e.g. *leukopenia*). Files containing concept names associated with the first two types of UMLS categories were considered as candidates for further consideration, and ones associated with the third type were eliminated. Files associated with the fourth type were examined further. A program was written to determine whether a concept name in each of those files had a matching lexical entry in MedLEE's lexicon. The semantic categories (as specified by MedLEE) of all concept names that matched were recorded in order to determine whether one MedLEE semantic category predominated for that particular UMLS category. If one did, it was considered the appropriate MedLEE category; if no MedLEE category predominated, the concept names in that file were eliminated because in order to generate the appropriate lexical entry for MedLEE manual review would be required.

The concept names in the remaining files were checked to see whether they had entries in the Specialist lexicon (not all words in the Meta are contained in Specialist). The Specialist lexicon contains the base form of a term as well as the variant forms, and use of Specialist provided a way to automatically identify the normalized target forms and the variant forms of the concept names. Finding a match required mapping Specialist entries to lower case letters. Additionally, the candidate concept names were also normalized by removing the symbol NOS, certain punctuation marks, and other symbols (i.e. <1>), and mapping letters to lower case. Concept names that were not found in Specialist were eliminated; the remaining names constituted the UMLS concept name list.

MedLEE English Modifiers: We observed that categories in MedLEE associated with general English modifiers (i.e. certainty types of information) had no relevant counterparts in the UMLS; concept names associated with these modifiers were assigned UMLS classes that were not useful for MLP purposes. For example *possible*, which generally occurs in reports as a modifier of a finding, was assigned the UMLS class **functional concept**; this category also contained the concept names *taxonomy* and *biosynthesis*. MedLEE requires a consistent and fine-grained classification of modifiers in order to determine the primary clinical information and associated modifier relationships in the text sentences. In order to obtain reasonable results for this study, we partitioned the MedLEE lexicon M-CUR into two components: an English modifier component and a clinically specialized component. The two lexicons LUMLS and M+UMLS both contained the English modifier component of M-CUR but utilized different clinical components.

Generating LUMLS: Lexicon LUMLS was created from the UMLS list of concept names in two stages, depending on whether the concepts were single- or multi-word phrases. Single word phrases were combined with the English component of the MedLEE lexicon, forming LUMLS'. To reduce the size of the lexicon multi-word phrases were only included if they were not comprised of phrases already contained in LUMLS'. For example, when an entry for *acanthocytosis* was added in the first stage, it was not necessary to add an entry for *hereditary acanthocytosis* in the second stage because MedLEE could handle that phrase if it occurred in a text report by treating *hereditary* separately as a modifier of *acanthocytosis*. To determine whether the multi-word phrases were covered by LUMLS', all multi-word phrases were treated as sentences and parsed using MedLEE and LUMLS'. Lexical entries were then generated only for those names that could not be parsed. LUMLS was created by augmenting LUMLS' with these new entries.

Generating M+UMLS: M+UMLS was constructed so that it added only those concept names in the UMLS term list that were not in M-CUR or that could not be parsed using M-CUR. First the single word concept names were checked so that names already in M-CUR were not considered further. A multiple stage process was also performed in order to reduce the size of the lexicon. Entries for single word concept names not found in M-CUR were first added to the entries of M-CUR forming M+UMLS''. The remaining multi-word concept names were then separated into two groups: those that were associated with body locations and those that were not. Those that were body locations were parsed using MedLEE and M+UMLS''; those concept names that could be parsed successfully were eliminated and the remaining body location concepts were then added to M+UMLS'' forming M+UMLS' (for example *accessory nerve* was added in this stage). The final step consisted of using MedLEE and M+UMLS' to parse the multi-word concept names that were not body locations; those names which could not be parsed successfully were added to M+UMLS' forming the final version M+UMLS. In the example shown above, the phrase *accessory nerve* occurred in 31 different UMLS concept names; once *accessory nerve* was added to the lexicon the other 30 phrases could be parsed by MedLEE (e.g. *neoplasm of accessory nerve*) and therefore were not included in M+UMLS.

Evaluating the Lexicons: We used the same training and test sets from the previous study, consisting of 40 and 79 cases respectively. In that study, queries were written by experts based on manual observation of output generated by MedLEE for the training cases. The MedLEE output consisted of target terms as specified by the MedLEE lexicon. For example, the target form for *CHF* was **congestive heart failure.** However, lexicons LUMLS and M+UMLS contained a number of lexical entries that were different from M-PRV (with possibly different target output forms) because they were based on the UMLS. For example, in LUMLS the normalized output for *CHF* was **CHF** and not **congestive heart failure** based on the UMLS. This meant that the original queries that were used to obtain the values for the variables had to be adapted. The training set of reports was processed using MedLEE and each of the new lexicons respectively, and the queries were subsequently modified by an expert based on the structured output generated.

In the prior study, a few of the variables were obtained using coded data, such as laboratory values. For this study we retained only those variables originating from textual discharge summary and chest x-ray reports. The reports in the test set were processed using MedLEE and lexicons M-CUR, LUMLS, and M+UMLS respectively, and the modified queries were executed to obtain final values for the variables. The values were compared with the reference standard. Sensitivity, accuracy, and specificity measurements were calculated for each of the variables, and the overall sensitivity, accuracy, and specificity values were computed by averaging the individual values.

RESULTS

We determined that only 31 of the UMLS semantic categories could be automatically mapped to a single category in MedLEE. There were 3,256 general English modifier types of lexical entries in MedLEE that were included in both lexicons generated for this study. LUMLS contained 67,405 entries that were obtained from the UMLS; 40,889 were single words. M+UMLS contained 14,630 entries from the clinical component of M-CUR, and an additional 50,846 entries from the UMLS; 40,501 were single words.

Eighteen variables were used in this study. Table 1 shows the results obtained by using MedLEE with the different lexicons to determine the values for the variables in the test set. M-PRV represents the results from the original study.

	M-PRV	M-CUR	LUMLS	M+UMLS
acc (cxr)	.96 (.94-.97)	.95 (.93-.97)	.88* (.85-.90)	.95 (.92-.97)
acc (dsum)	.91 (.89-.93)	.91 (.90-.93)	.83* (.81-.86)	.91 (..90-.93)
sens (cxr)	.73 (.59-.85)	.73 (.59-.85)	.50* (.38-.61)	.73 (.59-.85)
sens dsum)	.89 (.78-.93)	.89 (.78-.94)	.70* (.62-.78)	.89 (.78-.94)
spec (cxr)	.97 (.94-.99)	.96 (.93-.98)	.99 (.98-1)	.96 (.93-.98)
spec dsum)	.89 (.85-.93)	.89 (.84-.92)	.93 (..89-.95)	.89 (.84-.92)

Table 1. Accuracy (**acc**), sensitivity (**sens**), and specificity (**spec**) with 95% confidence intervals comparing variables using a reference standard determined manually by a clinical expert against ones obtained automatically by processing chest x-ray reports (**cxr**) and discharge summaries (**dsum**) using MedLEE and lexicons M-PRV, M-CUR, LUMLS, M+UMLS. Values followed by '*' signify values that were significantly different from values in corresponding columns (i.e. other lexicons). No other values were significantly different from the others.

Performance measures using M-PRV, M-CUR, and M+UMLS were not significantly different from each other, but the sensitivity and accuracy measures for LUMLS, which was created primarily using the UMLS, were significantly lower than all the other versions containing the MedLEE lexicon; it appears that LUMLS would be different from the others even with a Bonferroni correction for multiple hypotheses (except for specificity).

DISCUSSION

The results showed that when using LUMLS (the version based primarily on the UMLS) MedLEE did poorly. It was interesting that, compared to M-CUR (the version consisting of the MedLEE lexicon alone), no improvement was detected when using M+UMLS (the version consisting of the MedLEE lexicon supplemented by the UMLS) although M+UMLS had much broader coverage. This was probably due to the fact that M-CUR already contained most clinical terms found in reports associated with the domains of chest x-ray and discharge summaries. Many of the additional entries included in M+UMLS to supplement M-CUR probably rarely occurred in the test set of reports. Although the straightforward approach of acquiring lexical knowledge from the UMLS did not work, possibly more complex methods will. However, the advantage of using the UMLS for lexical knowledge was that it did not require medical expertise or the huge manual effort that was required to build the MedLEE lexicon.

We analyzed the errors that occurred using LUMLS and observed four major causes. One type of error was due to inadequate semantic classification in the UMLS for MLP purposes. For example, the UMLS semantic category **finding** corresponded predominately to the MedLEE category **finding**, and all concept names assigned that category were included in LUMLS; but the UMLS category also was associated with some concepts that were categorized differently in MedLEE because they were observed to have different linguistic distributional properties. For example in the UMLS the concept names *very*, *atrial fibrillation*, and *pulse rate* all are assigned the semantic class **finding**. However, *very* has a well-defined meaning only when it modifies a concept name, and it does not occur by itself in clinical reports; *atrial fibrillation* has a well-defined meaning and can occur by itself; *pulse rate* has a well-defined meaning and typically occurs in reports with a numeric value or with information denoting a change in value such as *increased.* In MedLEE these three concept names have different semantic categories **degree, finding,** and **bodymeas** respectively, enabling MedLEE to obtain a more accurate analysis by using rules that reflect the linguistic differences noted above.

A second source of error was due to the method we used for acquiring the lexical knowledge. Some relevant concept names were not included because they were not found in the Specialist Lexicon. For example, *shortness of breath* was in the Meta but not in Specialist. The method we used could be modified in the future to include all concept names in the Meta. However, then the lexicon would be huge, and would contain a large number of concept names that never appear in clinical reports. A method to remove highly unlikely concept names would have to be developed as well as a procedure to determine normalized forms.

Another primary source of error was due to lack of coverage by the Meta or to our exclusion of concept names in semantic classes that were determined to be inadequate for MLP purposes. For example, *infiltrate* was classified as a **functional concept**, which was determined to be an inadequate UMLS class. This problem was troublesome because concept names associated with those categories could not be acquired automatically. The problem associated with lack of coverage however may be addressed in subsequent versions of the UMLS. Until then, the only way to correct for this type of problem would be to use a manually created supplementary lexicon or possibly other terminological resources.

A fourth cause of error was due to the frequent occurrence of ambiguous abbreviations in the Meta. For example, *be* typically occurs in clinical reports as a common English verb but in the UMLS Meta *be* is an abbreviation that corresponds to the concepts **bacterial endocarditis, barium enema,** and **beryllium**. When using LUMS this caused MedLEE to interpret *be* incorrectly as a clinical term instead of as a verb. Resolution of the correct sense of an ambiguous term is a complex computational problem, which is discussed in more detail by Liu and colleagues[9].

Performance using M-PRV and M-CUR was not significantly different from each other; this signifies that changes made to MedLEE subsequent to the previous study did not change the outcome. Since MedLEE has undergone extension, it is important to verify that performance did not deteriorate because of extension to broader domains.

The performance measures reported in Table 1 show different results when using M-PRV than was previously reported on by Friedman[6]. This was due to two factors: eight variables associated with coded values that were used in the previous study were not used in the current study, and the method used for computing performance differed because this time we calculated performance measures for each variable independently and then averaged the results while previously the measures were pooled for the different variables.

SUMMARY AND CONCLUSIONS

We generated two lexicons by using the UMLS to automatically create lexical entries for an MLP system called MedLEE. We evaluated performance of MedLEE using the current MedLEE lexicon, and each of the two new lexicons. We found that the MedLEE lexicon alone performed no differently than the one that was supplemented by the UMLS. However, the one that was created primarily using the UMLS performed significantly worse. We also found the UMLS to be a valuable resource for MLP purposes because it substantially reduced the effort. We believe performance may be increased by improving the methods that generated the lexical entries.

Acknowledgement: This publication was supported in part by grants LM0624 from the National Library of Medicine and ILS-9817434 from the National Science Foundation.

References

1 Lindberg D, Humphreys B, McCray AT. The Unified Medical Language System. Meth Inform Med 1993; 32:281-291.

2 McCray AT, Srinivasan S. Automated access to a large medical dictionary: online assistance for research and application in natural language processing. Comput Biomed Res 1990; 23:179-198.

3 Baud RH, Lovis C, Rassinoux AM, Michel PA, Scherrer JR. Automatic extraction of linguistic knowledge from an international classification. In MEDINFO 98. 1998; 581-585.

4 Zweigenbaum P, Courtois P. Acquisition of lexical resources from SNOMED for medical language processing. In MEDINFO 98. 1998; 586-590.

5 Johnson SB. A semantic lexicon for medical language processing. JAMIA:1999; 6(3):205-218.

6 Friedman C. A broad coverage natural language processing system. In Overhage M, ed. Proc AMIA Symp 2000: 270-274.

7 McCray AT. The UMLS Semantic Network. Proc 13th Annual SCAMC. 1989; 503-507.

8 Friedman C, Knirsch CA, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp. 1999; 256-260.

9 Liu H, Lussier Y, Friedman C. A study of abbreviations in the UMLS. Proc AMIA. 2001 (in press)