

Comparing Metamap to MGrep as a Tool for Mapping Free Text to Formal Medical Lexicons

Samuel Alan Stewart*¹, Maia Elizabeth von Maltzahn², Syed Sibte Raza Abidi¹

¹NICHE Research Group, Dalhousie University, 6050 University Ave., Halifax, NS, Canada.
<http://www.cs.dal.ca/~niche>.

²Department of Internal Medicine, University of Saskatchewan, 103 Hospital Drive, Saskatoon, SK, Canada

*Contact Author: sam.stewart@dal.ca

Abstract. Metamap and Mgrep are natural language processing tools for mapping medical free text to formal medical lexicons, but an indepth comparison of the programs and their application to social media data has never been pursued. This project is interested in comparing the programs, in order to determine which program is most appropriate for mapping web 2.0 communication data. The archives of the Pediatric Pain Mailing List (PPML) were mapped with both programs, and each term returned was checked for correctness. The analysis resulted in Mgrep having a significantly higher precision (76.1% to 58.8%, difference of 18%, p-value < 0.0001) while Metamap returned more terms: 2381 to 1350. When considering only perfect or multiple matches, Mgrep still had better precision (81.2% to 71.3%, difference 10%, p-value < 0.0001). Ultimately Mgrep's precision may make it the better choice for many applications, but when there is more value in number of correct terms returned over accuracy of those terms, Metamap's larger set and superior scoring function may make it the tool of choice.

Keywords: Natural Language Processing; Semantic Mapping; MeSH; UMLS; Knowledge Management; Knowledge Translation

1 Introduction

Web 2.0 tools provide a valuable service to the healthcare community. Through online discussion forums, mailing lists, blogs, etc., clinicians can find mediums through which they can communicate their problems and share their experiences, developing relationships and creating a virtual community of practice (Wenger, 2004). Notwithstanding the evidence-based nature of modern healthcare, these online tools provide avenues for sharing experiential and tacit knowledge (Abidi, 2006) with colleagues in a way that spans the temporal and geographical boundaries that often prevent face-to-face communication.

The archives of these online conversations contain vast amounts of tacit and experiential knowledge. Extracting this knowledge and making it available to the community can improve the overall knowledge base, but how best to process this unstructured free text has proven a challenge.

Natural language processing approaches have been pursued in the past, including the semantic mapping of the unstructured text from the online tools to keywords from structured medical lexicons, such as UMLS (UMLS, 2012) and MeSH (MeSH, 2010). Of all the approaches to this mapping, the two most successful have been the Metamap program (Aronson, 2001) developed at the NLM, and Mgrep, the mapping tool of choice for the Open Biomedical Annotator (Jonquet et al., 2009).

These two programs take different approaches to the mapping process, and as such result in different sets of keywords when mapping the same source text. Previous research (Shah et al., 2009) has investigated comparing the two programs with respect to mapping the metadata associated with free, online databases, but this comparison did not explore the successes and failures of each program in any great detail, and the nature of metadata is very different from the archives of social media tools.

This paper is interested in comparing the results of using Metamap and Mgrep to map the archives of an unstructured medical mailing list to the MeSH medical lexicon. We first want to investigate general precision, to determine which program is more accurate with its mapping. We also want to delve deeper into the precision of the two programs, to determine if there is a relationship between mapping score and correctness, and we want to look at the overlap between the terms returned from the two programs.

The paper will proceed as follows: the background section will summarize the medical lexicon system, and the MeSH system in particular. It will explore some previous semantic mapping techniques, along with in depth explanations of how Metamap and Mgrep work. The methods section will outline the data preparation, the mapping process, and the analysis plan. The results section will summarize the analysis of the mappings by the two programs, and finally the discussion and conclusion sections will attempt to synthesize the analysis into a useful comparison of the two programs.

2 Background

In an evidence-based medical world, it is vital that knowledge be available to clinicians at the point of care. Unfortunately, the lack of organization, proper indexing, aging information sources and poor distribution have been shown to negatively affect a clinician's access to pertinent information (Covell et al., 1985; Timpka et al., 1989; Osheroff et al., 1991). The use of formal medical lexicons is a key step in improving clinician access to medical knowledge by providing a unified indexing of the existing medical knowledge.

The Unified Medical Language System (UMLS) is developed by the National Library of Medicine (NLM) to facilitate the computerization of medical knowledge, with the ultimate goal of allowing computer-systems to “understand” the meaning of biomedical and health text (UMLS, 2012). To this end they have created a number of tools, one of which is the “Metathesaurus”, a formal lexicon that is the aggregate of over 150 different medical lexicons. The Metathesaurus includes a semantic network, assigning each term in the UMLS to one of the 135 generalized semantic types, which in turn have 54 relations between them. For a full listing of the UMLS Semantic Types, visit http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html.

The Medical Subject Headings (MeSH) lexicon is one of the subsets of the UMLS (MeSH, 2010). MeSH is the NLM's own controlled vocabulary, and is used to index the MEDLINE database. There are 26,142 terms in the 2011 edition of MeSH, arranged in a hierarchical fashion descending from 16 independent root nodes.

The UMLS and MeSH provide a valuable indexing resource to the medical profession, but clinicians need to be able to leverage these semantic languages in order to make full use of the formal indexing. Leroy and Chen (Leroy and Chen, 2001) developed a system that processes general medical queries and returns a set of medical keywords from the UMLS. Cimino et al (Cimino et al., 1993) designed a system that maps clinician queries to a set of generic queries based on UMLS keywords. Both of these systems take questions and map them to formal terms from a medical lexicon, which, though a first step, is different from mapping unstructured free text to a medical lexicon.

2.1 Semantic Mapping Techniques

The process of mapping free text to formal medical lexicons (and specifically to the UMLS) has long been an objective of the medical research community. The value of having formal representation of ideas combined with the challenge of performing the task manually has made research into automated approaches very valuable. This problem is often linked to MEDLINE, which is manually indexed by MeSH terms (MeSH, 2010), and thus provides an objective reason to connect text to UMLS terms. MicroMeSH (Lowe, 1987) was one of the first attempts to do this, by providing a simple system to expand search queries to MEDLINE and provide a tool where users could browse the MeSH tree around the terms they searched.

CHARTLINE (Miller et al., 1992) processed free text of medical records and connected them to relevant terms in the MeSH lexicon via a direct mapping. This process was improved by SAPHIRE (Hersh and Greenes, 1990), which explored the idea of processing free text and cleaning it by mapping terms to their synonyms. This was a valuable addition to the literature, as it normalized the process of mapping *women* to *woman*. This process was taken up by Nadkarni et al (Nadkarni et al., 2001) who used this synonym mapping along with a part of speech tagger to better identify the structure of the conversations and attempt to identify specific words and phrases in the text. PhraseX (Srinivasan et al., 2002) also used this kind of synonym parser to analyze the mapping of MEDLINE abstracts to the UMLS metathesaurus, in order to evaluate the contents of UMLS itself. Other, similar approaches include KnowledgeMap (Denny et al., 2003) and IndexFinder (Zou et al., 2003).

The current, gold standard is Metamap, though another product, called Mgrep (Shah et al., 2009) provides a very similar service. The creators of the Open Biomedical Annotator (OBA) (Jonquet et al., 2009) designed a system that leverages the results of any semantic mapping service (Metamap or Mgrep) and the ontology relations within the lexicon to produce a more complete semantic mapping. The OBA authors decided to make Mgrep their default mapping service, due largely to its vastly quicker processing times, but their approach would work with Metamap as well.

2.2 Metamap

Metamap uses a special natural language parser called SPECIALIST (Aronson, 2001) to find all the nouns and noun-phrases in a discussion thread, and maps them to one or more UMLS terms. Each mapped UMLS term is assigned a score that is a measure of how strongly the actual term mapped to the UMLS vocabulary. The score is a weighted average of four metrics measuring the strength of the matching, with an overall range in [0,1000], with higher scores indicating a better match. The formal equation for calculating the scores is:

$$\frac{1000 \times (Centrality + Variation + 2 \times Coverage + 2 \times Cohesiveness)}{6} \quad (1)$$

- Centrality: An indicator of whether the matched (source) term is the head of the phrase
- Variation: A measure of the distance between the matched term and the root word. For example, if the source word is eye and the match is to the term ocular, the distance is 2, as ocular is a synonym for eye
- Coverage and Cohesiveness: Measures of how well the source term and the UMLS term match each other: if the source and UMLS terms are both “pain” then the match is perfect, but if the source term ocular matches to the UMLS term Ocular Vision then the coverage and cohesiveness are less than perfect.

Metamap’s precision and recall in previous projects have varied depending on the format of the text being processed, from values as high as 0.897 and 0.930 respectively (Kahn and Rubin, 2009) to values as low as 0.56 and 0.72 (Chapman et al., 2004). The difference between the precision and recall values show that Metamap does a good job at returning pertinent MeSH terms, but also returns impertinent terms as well, i.e., its results are somewhat noisy. Projects that reported low recall and precision with Metamap acknowledged that many of the problems come from the inherently ambiguous nature of the text being processed: in processing medical residents’ voice recordings, it was noted that Metamap failed to recognize abbreviations, acronyms or complex phrases that omitted key terms (Chase et al., 2009).

For our purposes, the Metamap scoring system provides a baseline measure of how well the mapped UMLS term represents the original term in the PPML discussion thread. Table 1 contains some sample mappings to the MeSH lexicon and their scores.

Despite the inconsistencies in the terms returned by Metamap, it provides a valuable tool for mapping unstructured messages and conversations to a structured medical lexicon. The Knowledge Linkage project (Stewart and Abidi, 2012) uses these mappings to try and provide explicit knowledge links to the experiential knowledge being shared within the community.

2.3 Open Biomedical Annotator and MGrep

The Open Biomedical Annotator (Jonquet et al., 2009) was developed to automate the process of providing keywords to datasets that are available on the web. Their process was to take the metadata from the datasets, pass them through a semantic mapping

engine (either Metamap or Mgrep) and then post-process their output using ontological relationships.

The authors of the Open Biomedical Annotator performed an experiment to compare MetaMap to Mgrep (Shah et al., 2009) in terms of accuracy and speed. They found that Mgrep performed slightly better in terms of precision and was much faster (1/5th of a second compared to 8 minutes). The authors concluded that, because they were looking for real-time implementation, Mgrep was a better option for them, and thus The Open Biomedical Annotator was implemented using Mgrep.

The details of how Mgrep works are not completely clear, and publications on it have been limited to conference posters (Dai et al., 2008). The authors of the Open Biomedical Annotator claim that it “implements a novel radix-tree-based data structure that enables fast and efficient matching of text against a set of dictionary terms” (Jonquet et al., 2009). The scoring algorithm as well is not completely explained, though it performs a similar expansion scoring to Metamap, where partial matches and derived matches receive lower scores than perfect matches. Mgrep is not distributed itself, but is accessed via the OBA: performing a mapping with the OBA without using the ontological expansions results in a strictly Mgrep-based mapping. Table 1 contains some sample mappings from Mgrep.

<i>The report stated that when music therapy is used, the babies required less pain medication. Does anyone know of any published reports of empirical research demonstrating the effect?</i>					
Metamap Terms			Mgrep Terms		
Source	MeSH Term	Score	Source	MeSH Term	Score
music therapy	Music Therapy	1000	Music	Music	10
			therapy	therapy	10
the babies	Infant	966			
less pain medication	Pain	660	Pain	Pain	10
less pain medication	Pharmaceutical Preparations	827			
of any published reports	Publishing	694	Report	Report	16
			Research	Research	10
of empirical research	Empirical Research	1000	Empirical Research	Empirical Research	10

Table 1: Sample message and its associated MeSH mappings from both Metamap and Mgrep

2.4 Conclusion

It is clear that Metamap and Mgrep are the two most popular options for mapping medical free text to structured medical lexicons. Minimal research has been done in

terms of comparisons, but more is needed, particularly within the mapping of social media data. Using MeSH as a target lexicon has the benefit of having many comparable projects, and the follow-up connection to MEDLINE and other sources that are indexed by MeSH is an additional reason to use it as a target lexicon.

3 Methods

The data for this project is the archives of the Pediatric Pain Mailing List (PPML) from January 2006 - December 2008. The data were originally extracted and processed for the Knowledge Linkages project (Stewart and Abidi, 2012) and the parsing and cleaning details are contained therein. For our purposes the content of the messages were extracted and cleaned to try and remove non-medical information (user signatures and reply-text being the major targets). An attempt was made to remove non-pertinent messages (such as conference announcements and job advertisements) as those types of messages do not contain the embedded medical knowledge that we are interested in. Once the data was cleaned and prepared it was mapped with both Metamap and the Open Biomedical Annotator (OBA), producing a set of terms and scores for each message from each program.

3.1 Mapping

In a paper by Abidi (Abidi et al., 2005) they outlined semantic filters they applied when using Metamap in mapping the content of clinical practice guidelines to formal medical terms. Of the 135 semantic types in the UMLS certain types, such as Amphibian or Professional Society, were not deemed pertinent to the subject, and were filtered out. 108 of the semantic types were used, while 27 were filtered out. The semantic types filtered out were: Amphibian, Animal, Bird, Class, Family Group, Fish, Functional Concept, Geographic Area, Group, Idea or Concept, Intellectual Product, Language, Mammal, Occupation or Disciple, Organization, Physical Object, Plant, Population Group, Professional Society, Professional or Organizational Group, Qualitative Concept, Quantitative Concept, Regulation or Law, Reptile, Research Device, Self-help or Relief Organization, Spatial Concept, Temporal Concept and Vertebrate.

The mapping was done using Metamap09. Though newer versions of Metamap have been made available the decision was made to use the same mappings that were done in the original project (Stewart and Abidi, 2012). Changes between versions of Metamap are minimal, so a change to the new version of the program is not expected to drastically affect the results.

For Mgrep, the mapping was done using the OBA REST services, available at <http://bioportal.bioontology.org/annotator>. The OBA has the same semantic type filters as Metamap, and the same filtering set was used. None of the OBA expansion options were used, resulting in the OBA returning a strictly Mgrep-mapped set.

In order to make the scores comparable between the programs, the Metamap scores were divided by 100, putting them on the same [0, 10] range as the Mgrep scores. For each program, the terms within a specific message were aggregated. This means that, though the range for an individual mapping score is [0,10], the scores can in reality go

from $[0, \infty]$, as there could be multiple mappings of the same term in a message. For the mappings reviewed, the maximum score returned was 128.26 for Metamap and 190 for Mgrep.

Once the mappings were created they needed to be checked. The messages and their mappings were reviewed by a medical expert. For each message the content of the message was first evaluated to determine if it was medically oriented, completing the filtering process that was somewhat handled in the data cleaning process. After that each MeSH term mapped to the message was reviewed and determined to be relevant to the conversation or not. The process was continued until 200 medically relevant messages had been found, with 127 messages being deemed not medically relevant.

3.2 Analysis

The analysis will begin with a simple investigation of the precision of both programs. Since both programs report scores for each mapping, an investigation of the relationship between score and correctness will also be investigated, to determine both the value of the scores being returned, and whether the scores could be used to improve the mapping process. We also want to compare the mappings between Mgrep and Metamap to study the overlap. The natural partner when studying precision is recall, but while precision, the proportion of returned terms that are correct, is relatively simple to calculate, recall, the number of correct terms that were found, is not nearly as simple to find, as this requires the correct terms for each of the messages to be pre-specified, which was not a feasible task for this project. Relative recall (Clarke and Willett, 1997) is often used to compare search strategies in which there is no annotated database to calculate recall from, but relative recall tends to favour system that return more results, and Metamap returned many more terms, and thus must have a higher relative recall. We will instead look at the overlap between the two programs and its relationship to precision.

4 Analysis

Table 2 presents some summary statistics for both Mgrep and Metamap. As we can see in the table, Mgrep had significantly higher precision, with a p-value < 0.0001 .

Program	# terms	# correct	Precision	difference	p-value
Metamap	2381	1384	58.12%		
Mgrep	1350	1027	76.07%	17.95% [14.9%,21.0%]	< 0.0001

Table 2: Summary of the mapping process for both programs. The p-value is calculated using a 2-sample z-test with a continuity correction.

4.1 Scores and Correctness

Though Mgrep has a higher general precision than Metamap, the relationship between score and correctness reveals that Metamap's precision may be better than it appears.

Figure 1 presents boxplots for both programs, comparing the scores for both programs between incorrect and correct mappings.

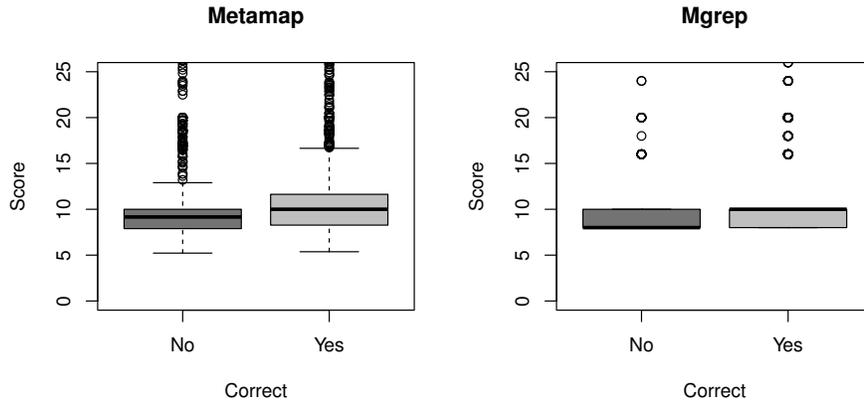


Fig. 1: Boxplots for comparing scores to correctness for both programs. Note that the plots are truncated to the [0,25] range for clarity.

For both programs it appears that there is a significant relationship between score and correctness, though the difference is more pronounced for the Metamap scores, as that program returns a wider range of scores. Infact, for individual terms Mgrep does not seem to return scores other than 8 or 10, with higher scores resulting from multiple mappings within the same message. Table 3 presents the comparison of correctness to score, and finds that, for both programs the correct terms have significantly higher scores.

		n	mean	Quantiles [5%,25%,50%,75%,95%]	Mean diff.	p-value
Metamap	Correct	1384	12.40	[6.38, 8.27, 10.00, 11.63, 28.27]	2.57	< 0.0001
	Incorrect	997	9.82	[5.94, 7.89, 9.16, 10.00, 19.01]		
Mgrep	Correct	1027	13.68	[8, 8, 10, 10, 30]	3.55	< 0.0001
	Incorrect	323	10.13	[8, 8, 8, 10, 17.8]		

Table 3: Comparing scores to correctness for both programs. The p-values are calculated using a Wilcoxon Rank-Sum test to account for the extreme skewness of the data.

The relationship between scores and correctness can be investigated further by looking at 10% quantiles of the data. Tables 4 and 5 report the correctness stratified by 10% quantiles of the scores. The quantiles of the Metamap scores are much more spread out, which is to be expected as their scoring algorithm is more complex, resulting in a

wider range of values. What is interesting, looking at the table, is that there seems to be a significant jump in precision for both programs for terms that score 10 points or higher. Table 6 looks at the relationship between correctness and score dichotomized to above/below 10 points.

<i>Quantile</i>	[5.22,6.6)	[6.6,7.55)	[7.55,8.61)	[8.61,8.75)	[8.75,9.28)	[9.28,10)	[10,18.6)	[18.6,128)
<i>Correct</i>	129	77	135	94	109	149	247	57
<i>Incorrect</i>	94	175	104	143	56	56	554	201
n	223	252	239	237	165	205	801	258
Precision	0.42	0.69	0.44	0.60	0.34	0.27	0.69	0.78

Table 4: Correctness by 10% quantiles of scores for Metamap. Note that quantiles that were the same were collapsed together, thus the quantile [10, 18.6) has 801 observations in it, which represents 3 quantiles of data.

<i>Quantiles</i>	[8,10)	[10,16)	[16,20)	[20,190)
<i>Correct</i>	162	126	19	16
<i>Incorrect</i>	328	445	69	184
n	490	571	88	200
Precision	0.67	0.78	0.78	0.92

Table 5: Correctness by 10% quantiles of scores for Mgrep. Because of the lack of range of Mgrep scores many of the quantiles were similar, and were thus collapsed into 4 groups from 10.

Metamap's precision has jumped from 58% to 71%, while Mgrep's has jumped from 76% to 81%. Though Mgrep's precision amongst only those terms that score ≥ 10 is still significantly higher (10% difference, 95% CI: [6.1%, 13.9%], p-value < 0.0001), Metamap improved its precision by 13%, whereas Mgrep only improved by 5%. It is clear that there is a significant relationship between score and correctness.

	Metamap Score		<i>Total</i>
	< 10	≥ 10	
<i>Correct</i>	628	756	1384
<i>Incorrect</i>	693	304	997
<i>Total</i>	1321	1060	2381
Precision	47.5%	71.3%	

	Mgrep Score		<i>Total</i>
	< 10	≥ 10	
<i>Correct</i>	328	699	1027
<i>Incorrect</i>	162	161	323
<i>Total</i>	490	860	1350
Precision	66.9%	81.2%	

Table 6: Looking at the relationship between score ≥ 10 and correctness for both programs.

4.2 Overlapping Terms

The overlap between the terms returned by Metamap and Mgrep presents an opportunity to try and evaluate the recall of the two programs. Though formal recall cannot be calculated, and relative recall is not valuable when one program returns so many more terms, studying what terms one program returned that another did not, and investigating what terms are missing, presents a valuable comparison of the two programs. Table 7 presents the overlap of the two programs with respect to correctness, and Figure 2 provides a visual representation of the difference.

<i>Program</i>	<i>Incorrect</i>	<i>Correct</i>	<i>Precision</i>	<i>Total</i>
Metamap Only	800	621	0.437	1421
Mgrep Only	126	264	0.677	390
Both Programs	207	782	0.791	989

Table 7: Comparing the overlap of the two programs. The precision reported is the number of terms for that row that are correct, i.e., it is the *Correct* column divided by the *Total* column.

The overlap of the two programs presents some interesting results. Of the 1350 terms returned by Mgrep, 989 were also returned by Metamap, resulting in an overlap of 73%. With 2381 terms returned, 41% of the terms returned by Metamap were also covered by Mgrep. Put in another way, if one were to only use Metamap, there would have been 264 correct mappings that were missed, while if one were to only use Mgrep there would be 621 correct mappings missed.

As demonstrated in Figure 2, the terms where the programs overlapped were more likely to be correct, with an overlap precision of 79.1%. This also leads to both programs having lower precision on the terms that only they returned than their overall average precision.

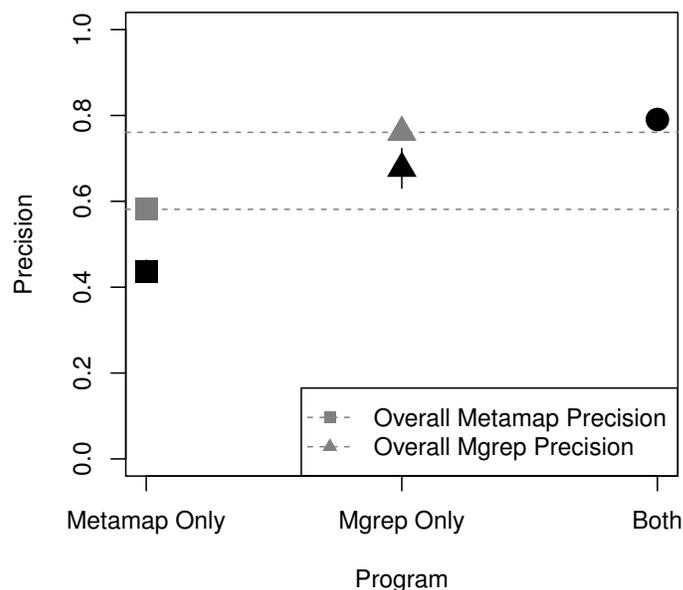


Fig. 2: Comparing the overlap of the two programs to their precision.

5 Discussion

Based strictly on precision, Mgrep outperforms Metamap. A difference of nearly 18% confirms the findings of (Shah et al., 2009) in their original investigation of the two programs. There is much more depth to the comparison, however, which reveals the potential utility of Metamap in certain situations.

Though both programs provide mapping scores, Metamap's seem more useful, providing both a wider range of scores and a larger difference in precision between the low and high scoring terms. One of the challenges of this comparison is a lack of details on how the Mgrep scoring algorithm works, but, though the authors claim a range of [0,10], in reality only 8's and 10's were returned (with higher scores all being aggregates of those two numbers).

Of particular interest is the poor performance of terms returned by Metamap that have scores just below perfect: Looking back at Table 4, the fifth decile, [8.75, 9.28), has a precision of only 34%. Looking into the mappings in this quantile, we see mappings that are based on variations in the root word, along with words that are based on a less than perfect coverage. The mappings in this group are inaccurate because they are taking a source term like "replacing" and mapping it to the MeSH term "Replantation",

which is rarely going to be the correct mapping. In an attempt to dig deeper into the potential variations on source terms, Metamap seems to be hurting its overall precision.

When mappings are restricted to only perfect matches (or less than perfect matches that occur multiple times), the precision of both programs increases, but the increase is more dramatic for Metamap (see Table 6). Previous studies that have investigated Metamap could improve their results by putting more effort into leveraging the Metamap scores.

This does not mean that terms that score less than perfect should necessarily be dropped, however, as there is a more to the evaluation of the two programs than precision. Looking back at table 6, removing all mappings with scores < 10 would drop 693 correct Metamap mappings and 162 correct Mgrep mappings. If the objective of the mapping process is strictly about precision then this may be a logical step, but if the objective is to try and find suggested terms to provide to the users, then there is little harm in providing incorrect suggestions, especially if it leads to more pertinent terms being provided as well. Looking at the overlap of the two programs, though Mgrep had a higher precision, it missed 621 terms that Metamap provided, terms which may have been beneficial to the user. Likewise, there are 264 terms missed by Metamap that were returned by Mgrep, which could also have been helpful.

If the objective of the mapping process is strictly to be as precise as possible, then using Mgrep and restricting the mapping solely to terms that score 10 points will result in the most accurate mapping. If you are developing a suggestion engine, however, or if your system can leverage the mappings scores, as our Knowledge Linkage program did (Stewart and Abidi, 2012), then perhaps the larger set returned by Metamap, combined with the superior scoring function, may be more useful to your project.

Though it was not studied formally in this project, we did find that Mgrep was vastly faster than Metamap, even when used over the internet through their REST services. This confirms the findings of (Shah et al., 2009), and if you are trying to develop a real-time system then Metamap may be too slow for your application.

6 Conclusion

There is an obvious need for indexing engines that can process free text and match them to formal medical lexicons. Though this project focused on MeSH, there are obvious expansions to any component of the UMLS, and mappings to ICD and SNOMED can provide valuable resources to those working in health information technology.

The mapping of social media archives to MeSH is a challenging objective. A precision of 58% by Metamap is at the low end of the range of precisions reported by other papers that studied program (Chapman et al., 2004; Chase et al., 2009), and the challenges of mapping abbreviations, acronyms and complex phrases from medical charts continue to be a problem for the mapping of social media data. This does not mean that the mapping process cannot be used, but when leveraging the terms provided by these programs the potential for incorrect mappings must be taken into account.

This project had some shortcomings. A double review of the mappings rather than a single review would have provided more confidence in the “correctness” of the mappings. The Metamap program used was the 2009 edition, as those were the mappings

that were produced for the Knowledge Linkage project (Stewart and Abidi, 2012), and there have been multiple releases since then. Re-running the analysis with the new program would probably not change the precision of Metamap significantly, but it would certainly change some of the mappings. We believe that the general structure of the analysis would remain the same, however a comparison of the old and new versions should be investigated. More details of how Mgrep works need to be made available, especially with respect to the scoring algorithm. As well, the aggregation of multiple mappings needs to be broken down, which could be used to expand the results in section 4.1. Correct/Incorrect may not be the best way to classify mappings: providing the term “Pain” in a discussion of needle stick injuries is not incorrect, but it is not as useful as the MeSH term “Needle Stick”. Re-evaluating each mapping on a 5-point Likert Scale may provide more valuable insights.

Developing a way to measure some form of recall would improve the analysis: studying the crossover between the two programs is helpful, but being able to identify and study what was missed is a valuable component of the comparison of the two programs. Each message could be reviewed, and the potential MeSH terms that are not present could be recorded, providing some insight into terms that were not mapped. This analysis will be done in future work.

Moving forward, the programs are best measured not by evaluating their correctness in terms returned, but by their utility embedded in other programs. Re-implementing the Knowledge Linkage project with Mgrep and re-running the analysis from that project (Stewart and Abidi, 2012) would be a stronger way to measure whether Mgrep is more or less useful in mapping free text to medical lexicons. A larger review set would also allow a more indepth analysis of the correctness as a function of position in the MeSH tree, both in terms of source root and depth from the top.

Bibliography

- Abidi, S. (2006). *Healthcare Knowledge Sharing: Purpose, Practices, and Prospects*, chapter 6, pages 65–86.
- Abidi, S., Kershaw, M., and Milios, E. (2005). Augmenting gem-encoded clinical practice guidelines with relevant best evidence autonomously retrieved from medline. *Health Informatics Journal*, 11(2):95–110.
- Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: The metamap program. *Proceedings of the AMIA Symposium*.
- Chapman, W. W., Fiszman, M., Dowling, J. N., Chapman, B. E., and Rindfleisch, T. C. (2004). Identifying respiratory findings in emergency department reports for bio-surveillance using metamap. *MEDINFO*.
- Chase, H. S., Kaufman, D. R., Johnson, S. B., and Mendonca, E. A. (2009). Voice capture of medical residents' clinical information needs during an inpatient rotation. *Journal of the American Medical Informatics Association*, 16:387–394.
- Cimino, J., Aguirre, A., Johnson, S., and Peng, P. (1993). Generic queries for meeting clinical information needs. *Bulletin of the Medical Library Association*, 81(2):195–206.
- Clarke, S. J. and Willett, P. (1997). Estimating the recall performance of web search engines. In *Aslib Proceedings*.
- Covell, D., Uman, G., and Manning, P. (1985). Information needs in the office practice: are they being met? *Annals of Internal Medicine*, 103(4):596–599.
- Dai, M., Shah, N., Xuan, W., Musen, M., Watson, S., Athey, B., and Meng, F. (2008). An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics*.
- Denny, J. C., Smithers, J. D., Miller, R. A., and Spickard, A. (2003). understanding medical school curriculum content using knowledgemap. *JAMIA*, 10:351–362.
- Hersh, H. and Greenes, R. (1990). Sapphire - an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Comput Biomed Res*, 23:410–425.
- Jonquet, C., Shah, N. H., and Musen, M. A. (2009). The open biomedical annotator. *Summit of Translational Bioinformatics*, pages 56–60.
- Kahn, C. E. J. and Rubin, D. L. (2009). Automated semantic indexing of figure captions to improve radiology image retrieval. *Journal of the American Medical Informatics Association*, 16:280–286.
- Leroy, G. and Chen, H. (2001). Meeting medical terminology needs—the ontology-enhanced medical concept mapper. *IEEE Transactions on Information Technology in Biomedicine*, 5(4):261–270.
- Lowe, H. (1987). Micromesh: a microcomputer system for searching and exploring the national library medicines medical subject headings (mesh) vocabulary. *Proc Annu Symp Comput Appl Med Care*, pages 717–20.
- MeSH (2010). Medical subject headings. <http://www.nlm.nih.gov/mesh/>.
- Miller, R. A., Gieszczykiewicz, F. M., Vries, J. K., and Cooper, G. F. (1992). Chart-line: Providing bibliographic references relevant to patient charts using the umls

- metathesaurus knowledge sources. *Proc Annual Symposium of Comput Appl Med Care*, pages 86–90.
- Nadkarni, P., Chen, R., and Brandt, C. (2001). Umls concept indexing for production databases: a feasibility study. *JAMIA*, 8:80–91.
- Osheroff, J., Forsythe, D., Buchanan, B., Bankowitz, R., Blumenfeld, B., and Miller, R. (1991). Physicians' information needs: analysis of questions posed during clinical teaching. *Annals of Internal Medicine*, 114(7):576–581.
- Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P., and Musen, M. A. (2009). Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10 (suppl 9):S14.
- Srinivasan, S., Rindfleisch, T. C., Hole, W. T., Aronson, A. R., and Mork, J. G. (2002). Finding umls metathesaurus concepts in medline. *Proc AMIA Symp*, pages 727–731.
- Stewart, S. A. and Abidi, S. S. R. (2012). An infobutton for web 2.0 clinical discussions: The knowledge linkage framework. *IEEE Transactions on Information Technology in Biomedicine*, 16(1):129–135.
- Timpka, T., Ekstrom, M., and Bjurulf, P. (1989). Information needs and information seeking behavior in primary health care. *Scandinavian Journal of Primary Health Care*, 7(2):105–109.
- UMLS (2012). Unified medical language system fact sheet. Web. <http://www.nlm.nih.gov/pubs/factsheets/umls.html>.
- Wenger, E. (2004). Knowledge management as a doughnut: Shaping your knowledge strategy through communities of practice. *Ivey Business Journal*, pages 1–8.
- Zou, Q., Chu, W. W., Morioka, C., Leazer, G. H., and Kangarloo, H. (2003). Indexfinder: A method of extracting key concepts from clinical texts for indexing. *AMIA Annu Symp Proc*, pages 763–767.