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

Failure Analysis of MetaMap Transfer (MMTx)

Guy Divita, Tony Tse, Laura Roth

Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA

Abstract

A pilot study was conducted to evaluate the performance of the MetaMap Transfer (MMTx), a tool that extracts terms from free text and suggests matches to concepts in the Unified Medical Language System® (UMLS®). Five participants, including a content domain expert and a UMLS Expert, manually extracted and mapped terms to UMLS concepts for two disease summary documents from NLM's consumer health site, Genetic Home Reference. The resulting adjudicated annotations were used as a gold standard. Differences in automated term extraction and mapping between MMTx and MetaMap were noted. A failure analysis was conducted to categorize the types of terms not correctly mapped by MMTx. The most frequent type of failure (30%) resulted from missing inferential or world knowledge. Characteristics of each category are discussed. We distinguish between classes of failures that may be easily rectified, such as alternative retrieval strategies to extract exact matches, and ones that require additional research, such as coordinating conjunctions, co-reference resolution, and word sense disambiguation.

Keywords:

Natural Language Processing, Unified Medical Language System, Information Storage and Retrieval, Evaluation Studies

Introduction

Since the release of MetaMap Transfer (MMTx)[1], the recall has often been questioned. Because of its potential widespread usage in various applications, we have begun an investigation of MMTx coverage. This work is a first approximation at identifying and classifying MMTx. failures Such knowledge will facilitate future development efforts for improving recall. Because of the difficulties assessing information accurately using traditional methods (e.g., gold standards and relevance judgments), our approach was to describe types of MMTx errors (i.e., causes of failure). In this paper, we describe a pilot study to investigate the performance of MMTx (default configuration) in identifying terms in medical text found by humans and in retrieving relevant concepts from the Unified Medical Language System® (UMLS®) Metathesaurus (MTH).

Background

The UMLS [1,2] MTH organizes concepts and terms from multiple terminologies in a systematic and meaningful way. The biomedical concepts in the MTH come from controlled vocabularies and classifications used in patient records, administrative health data, bibliographic databases, and full text databases. Terms strongly associated with one medical meaning are aggregated into a set of synonymous or closely related forms, known as an MTH concept. Each is annotated with semantic types from a Semantic Network that broadly covers the medical domain. The meaning of a concept is derived from the semantic type, definition (if available) and location of component synonyms in their source hierarchies.

MetaMap [3] is a program that maps text to MTH concepts. It segments text into phrases, linking them to the closest matching concept(s) by exploiting synonymy, an integral part of the MTH. MetaMap also exploits lexical information from the SPECIALIST Lexicon and utilities [4] to generate lexical variants to find candidate mappings to the MTH. These mappings add computable semantics to the text. MetaMap Transfer (MMTx) is a Java implementation of MetaMap.

Related Work

Recently, Denny et al. [5] compared the KnowledgeMap (KM) concept identifier system to MetaMap. Manually annotated terms from medical school lecture documents served as a gold standard. Both KM and MetaMap use components of the UMLS (e.g., SPECIALIST lexicon, MTH), but each implements system-specific linguistics heuristics and scoring algorithms for candidate concepts. KM, also takes advantage of genre-specific style, structure, and format. KM includes an intriguing technique to ameliorate a source of failure: coordinating conjunctions. Nadkarni, Chen, and Brandt [6], in a feasibility study to determine if automatic concept indexing could be precise enough for a production system, identified seven specific failure categorizations. Although their system did not include a hand-tagged gold standard, it did include a good characterization of the source of their system's failures. Our pilot study combines a comparison using a gold standard with an in-depth failure analysis. Others have also explored the performance of MetaMap [7] and MMTx [8] on different text genres.

1. http://mmtx.nlm.nih.gov

Methods

We wanted to compare MMTx with complete and well-annotated text. No prior test collection that suited our purpose was found, so we explored the feasibility of creating our own test collection. We conducted a preliminary study to gain a better appreciation for the types of problems we were likely to encounter during the manual tagging process. We manually tagged several consumer health texts on diabetes and allergies. However, we had no domain experts to help us with UMLS or content issues that arose. We ran the texts through MMTx, compared the results and used the resulting insights to prepare guidelines to improve consistency in tagging among individuals.

The genetic condition documents from the NLM's Genetics Home Reference (GHR) web site [9] were chosen for this pilot study. The GHR provides summary information about genetic conditions and related genes, written for laypersons. We chose the GHR, in part, because the content providers were available for consultation when we set out to do the manual annotations.

Five participants were involved in the annotation process. Although direct access to the author of the documents would have facilitated determining intended meanings where ambiguous, a GHR document content editor participated in the annotation process. While none of the participants had complete knowledge of the over 100 UMLS component vocabularies to confirm term-MTH concept mappings definitively, someone involved with MTH editing and development (Laura Roth) was also a participant. In addition, a linguist and two informed consumers with some knowledge of the UMLS, (Guy Divita, Tony Tse), participated in the study.

All of the participants annotated each document. A combination of adjudication and consensus was used to choose final annotations through a series of meetings. Disagreements over competing MTH concepts were resolved by the manager of the MTH editors, while disagreements over content (i.e., intended meaning of a particular term or phrase) were resolved by the GHR content editor. This was a time-consuming process, and consequently, because of time constraints, only two documents, Infantile-onset Ascending Hereditary Spastic Paralysis[10] and Retinoblastoma [11], were completely analyzed for this task. These documents were chosen at random from the 100 documents on-line at the time.

The documents were preprocessed by the tokenizer component of the SPECIALIST NLP Tools[1] to break the text into tokens and listed, one per line, along with a token number and character offsets. The tokens were imported into Microsoft" Excel spreadsheets, where the annotators added bracketing, concept names, and concept unique identifiers (CUI's).

The annotators were asked to identify medical terms via bracketing. Our definition of medical terms consisted of the loose notion of "something that would be of interest to a medical professional". The annotators were asked to find the closest matching UMLS 2003AA concept for each bracketed term within the text. They were to map to the most specific concept possible. Although the intention was to identify one concept per term, multiple concepts were necessary to cover the bracketed term in certain instances. The annotators identified referential phrases with the referent concept and noted "co-reference".

Although the guidelines developed during the preliminary study were distributed, they proved to be insufficient, as an overarching guiding purpose, for these annotations had not been defined (e.g., information retrieval versus inferencing). Depending on the purpose, the text may be annotated differently. For example, composite noun phrases would be selected to support an information retrieval search capability, whereas atomic phrases would be selected to support logic processing or inferencing capabilities. It was subsequently agreed to annotate as if one were to support an information retrieval search capability, looking for longer pre-coordinated matches within the MTH to cover medically interesting terms rather than multiple concepts that covered the same surface terms.

The annotators were allowed to use any means possible to find matches, as long as they did not use MetaMap, MMTx, or the approximate matching facility within the Knowledge Source Server. They did use the basic normalized string concept retrieval, the normalized word focused searches, and the semantic navigator services of the Knowledge Source Server http://umlsks.nlm.nih.gov. A NLM internal browser from the MTH editing environment also proved useful. The annotators were encouraged to review the definition, semantic locality, and hierarchical context of each candidate concept, even when the surface form was an exact match, to ensure that the appropriate term was considered.

Each document was run through MMTx, Version 2.3, using the 2003[AA] UMLS MTH. The default MMTx options were employed. Note that MMTx, in its default configuration, does not use a part of speech tagger.[2]

A program was written using the MMTxAPI to print out the top mapping (best covering concepts of a phrase) in a format that aligned well with the hand-annotated document. The output was subsequently imported into Excel.

Although the intention is to develop an automated mechanism to evaluate MMTx with the hand annotations, this was done by hand for the pilot study by putting the two spreadsheets side by side to compare them. For this study, the top ranking mapping was compared with the human annotation. Only exact CUI matches were considered matches.

This pilot study also served to see if there were other un-intended differences between MetaMap and MMTx. To facilitate this, the same documents were run through the current version of MeMap, using the 2003 UMLS dataset. It is the intention of the MMTx developers to stay in sync with MetaMap.

1. The SPECIALIST NLP Tools are the section/sentence/phrase/term word tokenizers embedded within MMTx and also available as an independent package.

2. MMTx has no integral part of speech (POS) tagger, but relies on an external POS tagger server when locally available, using a tagger client *interface*. Others can use a locally trained tagger to integrate that capability into MMTx.

Results

Table 1: Token and Word Counts of the Documents

Document	Spastic Paralysis	Retino-blastoma
Tokens (ıncludes counts of punct)	850	916
Words (from Unıx tool "wc")	733	801

Table 2: MMTx Results

Total number of tokens within the documents	1766
Total number of terms [bracketed phrases] manually identified	316
Total number of UMLS concepts manually identified	314
Total number of manually identified UMLS concepts that MMTx found	152

Out of 316 terms identified as medical terms within the text, MMTx found 169 concepts, or about 53%. This number should not be construed as MMTx's recall for several reasons. This pilot study only included two documents. The recall is hypothesized to be sensitive to the style, form, and readability. The recall is also dependent on how much coverage the MTH has within the domain. The concept coverage of the two documents annotated was very good, with only two of the 316 medical terms identified missing suitable concepts from the UMLS.

Failure Analysis

Thirteen types of failure were identified (Table 3). Several cases had multiple points of failure, resulting in categorized totals that are more than the total number of missed terms.

Implicit meaning

This was the most prevalent source of failures, caused by inferred or contextual knowledge required to map a term to a concept when the information is not explicit in the surface form of the phrase in question. For example, within the context of retinoblastoma, an eye disorder, the phrase *family history of the disease* was mapped to the concept "FH: Eye disorder NOS, C0455396". In other cases, a term was mapped to a more specific concept due to inferences that were made. For example, the term *deletions* (e.g., "Testing is available for deletions of chromosome 13") was not mapped to "Gene Deletions, C0017260", which includes *deletions* as a synonymous string, but was mapped to "Chromosome Deletion, C0008628".

Narrower term not in MTH

Narrower terms missing from the MTH resulted in annotators choosing a broader concept because a specific form was not in the MTH, and the broader term had little or no lexical resemblance to the identified term. For example, *Infantile-onset ascending hereditary spastic paralysis* was mapped to "Spastic Paraplegia, Hereditary, Autosomal Recessive, C0751603". This happened to be both the title and the most prevalent term in this document, inflating the frequency of this type of error within this study.

Definitional phrase

This is a subcategory of the *meaning is implicit in the document* type of failure. It happened frequently enough to note it as a separate category. Text that failed to match terms from this category included phrases that were definitions or explanations of a term, often explicitly referencing the term afterwards. For example, the phrase *part of the eye that detects light and color* was annotated with the MTH concept "Retina, C0035298" from the sentence fragment *from the retina, which is the part of the eye that detects light and color*.

Table 3: Types of Missed Matches within MMTx

Category of Missed Match	Count	%
Implicit meaning	45	30%
Narrower term not in MTH	20	13%
Definitional phrase	15	10%
Co-reference	13	9%
Coordinating conjunction	10	7%
Bugs within MMTx	10	7%
Word sense ambiguity	8	5%
Missing synonym, concept in MTH	7	5%
X_form of_Y = X_Y	6	4%
Split phrase	6	4%
Missing concept in MTH	6	4%
Missed exact matches	2	1%
Missed synonymy in MTH	2	1%
Total	150	100%

Co-reference

This was a frequent source of mismatches. For example, *this disease* in the retinoblastoma document was mapped to "Retinoblastoma, C0035335" when the phrase *this disease* was seen in the text referring to the term *retinoblastoma* from the previous sentence.

Coordinating conjunctions

Coordinating conjunctions or complex syntax caused failures where phrases contained an abbreviatory way of stating multiple thoughts. Examples include *hereditary or sporadic form of retinoblastoma*, which should have matched "Hereditary Retinoblastoma, C0751483" and "Sporadic Retinoblastoma, C0751484". In some cases, it was difficult to interpret the coordination. For example, *progressive muscle weakness of the arms, legs and facial muscles* could mean "weakness of the arms, weakness of the legs, and weakness of the facial muscles" or "weakness of the arm muscles, weakness of the leg muscles, and weakness of the facial muscles". The annotators chose the broader concept "MUSCLE WEAKNESS, PROGRESSIVE, C0240421".

Bugs within MMTx

MMTx bugs were uncovered. For example, MMTx handles parenthetical expressions incorrectly.

Word sense ambiguity

The MTH contains concepts that share the same surface form, but not the same meaning. MMTx suggested both senses, ranked them the same, and chose the incorrect sense. For example, for the term *methods*, MMTx returned the concept "Methodology, C0025664" for the sentence fragment *therapists can recommend*

methods and devices, but the annotators chose "Methods, C0025663" as the appropriate term for this context.

Missing synonym, concept in MTH

These were cases where MMTx missed the match because it used a near synonymous form that was not present in the MTH, but where the concept was otherwise represented within the MTH. For example, the term *alsin* was missed when used in the context of the phrase, *a protein called alsin*. The annotators noted that this should have matched to the concept ″Alsin, Human, C1098466″. Likewise, the phrase *cell membrane organization* was missed, but the concept "Cell Membrane Structures, C0887867″ was used to cover that phrase.

X_form_of_Y = X_Y

Patterns of the nature X_**form_of**_Y were identified as medical terms. Further, they were missed by MMTx. Upon further investigation, terms that followed the pattern often were found within the MTH as X_Y. For instance, *The hereditary form of retinoblastoma* within the text was missed by MMTx, but was annotated with the concept "Hereditary Retinoblastoma, C0751483". Similarly, patterns of the nature X_**of**_Y were often identified as medical terms but missed by MMTx. Terms of this pattern often were annotated with concepts that were expressed as Y_X. *Images of the brain* was missed by MMTx, but annotated with the concept "Brain Imaging, C0203860".

Split phrase

These were the categorization of two common types of failures. The first includes cases in which a term was assigned an incorrect part of speech tag and consequently, split into two phrases. In the example *improved quality of life*, the word improved was mistagged as a verb, and consequently, *improved* was considered a separate phrase. When the POS tagger was employed, this word was identified as an adjective and kept with the phrase. This failure category also includes phrases split on prepositions. In the prior example, *of life* was segmented into a separate phrase. Both MetaMap and MMTx have options to combine these composite phrases to help alleviate failures of this sort. This option could be turned on to combine phrases into larger composite phrases to retrieve these concepts.

Missing concept in MTH

These were cases where the concept was not in the MTH. The annotators could not find a good representative concept that meant *ALS2 gene* or *symptom relief*.

Missed exact matches

These were cases where an exact surface form was the appropriate concept annotated but both MetaMap and MMTx missed it. Two examples include *both eyes* and *family history of*. Both failures were caused by how prepositions are handled.

Missed synonymy in MTH

Missed synonymy results in failures due to word sense disambiguation on MMTx's side and where the two senses of the term really could have been one sense. For example, the phrase *poor vision* was annotated with the concept "Poor, Vision, C0848430″, but MMTx chose "Vision, Low, C0042798″. These appear to be very similar terms, based on all their respective attributes.

Differences were found between MMTx and MetaMap

There were seven instances where the use of the tagger created differences when MMTx was compared against MetaMap. Not all of these differences were beneficial, as in the sentence *patients also experience slow eye movements* when MetaMap broke the phrase into the words *experience*, *slow*, *eye* and *movements*, missing the term *eye movements*. MMTx, however, did not break on *experience*, and erroneously left this sentence as one phrase. Neither MMTx nor MetaMap mapped the slow eye movements to ″Abnormal Ocular Motility, C0497202", the concept noted by the annotators. As noted above, it was found that MMTx does not handle parenthetical expressions correctly, attaching them to the phrase to the left of them, rather than splitting them into a separate phrase. It was also found that the ordering, in which candidates are returned, when the candidates are of the same ranking, is different. It was found that MMTx generates some poor variants and retrieves poor concepts for those variants whereas MetaMap does not. MetaMap includes a stop phrase feature that does not process phrases known to produce no results. An example of this occurred when MMTx found the concept "Helplessness, C0814060" for the term *help*, but MetaMap did not. These issues will be resolved during the next development cycle of MMTx.

Discussion

Some classes of these failures should and could be rectified. Other pattern matching techniques may retrieve exact matches that we have up until this time missed, either because they spanned multiple phrases or because they contained stop words. The X_**form**_Y and similar patterns may offer another way of suggesting variants to retrieve concepts. The quick composite option should help to retrieve terms that span multiple phrases. Of course, if problems such as the parenthetical expression tokenization bug were addressed, a fair number of concepts would also be picked up.

There are classes of failures that will require additional research to address, but are not beyond amelioration. A fair number of failures were caused by co-reference. There is a growing body of research that now addresses co-reference resolution. Coordinating conjunctions, like co-reference resolution, is an issue that will require additional research. Denny, Smithers, Miller and Spickard (2003) offer one approach. Word sense disambiguation efforts currently underway should also be looked at to ameliorate failures caused by multiple senses.

Some failures will be addressed over time with an evolving MTH. The classes of failure that are due to MTH content issues include missed synonymy in the MTH, synonyms missing where the concept was represented, more specific terms not found in the MTH, and finally concepts missing from the MTH.

Those failures that are due to inferential associations between the phrase and their associated concept appear to be much less tractable.

The failures thus far found have less to do with the particular idiosyncrasies of the documents mapped and more to do with the

mapping methods employed, the MTH, and the complexities of written language. As such, these issues will exist regardless of document genre, though not necessarily in the proportions reported here.

Future Work

The bugs and synchronization issues between MMTx and MetaMap will be addressed. Additional hand annotation efforts are envisioned to continue to draw useful strategies for better retrieval. In a future iteration, a semi-automated annotation environment customized to retrieve candidate UMLS concepts would make hand annotation a much more tractable task. This pilot study's comparison was done manually, but subsequent evaluations should be done by an automated method for evaluating hand-annotated documents with MMTx. Future evaluations should compare MMTx with options that optimize the performance for that task. For IR tasks, quick composite phrases should be turned on.

Those failure types that look to be tractable will be further investigated, including some mechanism to retrieve exact matches that are otherwise stopped out, or that span multiple phrases, and the pattern driven variant generation.

Conclusion

In this pilot study, we identified and classified 13 categories of manually reviewed term-concept matches that MMTx does not currently detect. Some of these failures may be easily rectified (e.g. Bugs within MMTx, Missed exact matches) while others require further research (e.g., Co-reference, Word sense ambiguity, Coordinating conjunction). One type (Implicit meaning) is clearly beyond the scope of MMTx. We also observed differences between MMTx and MetaMap, which will be corrected. Furthermore, the insights gained from developing the failure analysis guidelines and methodology will inform implementation of annotation support tools and continued formative evaluation of MMT.

Acknowledgments

Special thanks go to Allen Browne and Diane Mucci for their participation in annotating the GHR texts.

References

[1] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993 Aug; 32 (4): 281-91.

[2] Humphreys BL, Lindberg DA. Building the Unified Medical Language System. Proc Annu Symp Comput Appl Med Care 1989; 475-80.

[3] Aronson, AR, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001; 17-21.

[4] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994; 235-9.

[5] Denny JC, Smithers JD, Miller RA, Spickard A 3rd. "Understanding" medical school curriculum content using KnowledgeMap. J Am Med Inform Assoc. 2003 Jul-Aug; 10(4):351-62. Epub 2003 Mar 28.

[6] Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. J Am Med Inform Assoc. 2001 Jan-Feb;8(1): 80-91.

[7] Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. J Biomed Inform. 2003 Aug-Oct;36(4-5):334-41.

[8] Pratt W, Yetisgen-Yildiz M. A study of biomedical concept identification: MetaMap vs. people. Proc AMIA Symp. 2003;:529-33.

[9] http://ghr.nlm.nih.gov

[10]http://www.ghr.nlm.nih.gov/ghr/disease/infantileonsetascendinghereditaryspasticparalysis

[11]http://www.ghr.nlm.nih.gov/ghr/disease/retinoblastoma

[12]http://umlsks.nlm.nih.gov

Address for correspondence

Guy Divita
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894
divita@nlm.nih.gov