Proceedings of AMIA Annual Symposium 2006:589-593.

Besides Precision & Recall: Exploring Alternative Approaches to Evaluating an Automatic Indexing Tool for MEDLINE

Aurélie Névéol[1, 2, 3], Ph.D., Kelly Zeng[1, 2], M.S., Olivier Bodenreider[1, 2], M.D., Ph.D.
[1]US National Library of Medicine, Bethesda, Maryland
[2]National Institutes of Health, Department of Health & Human Services
[3]Équipe CISMeF and GCSIS, INSA & Université de Rouen, France
{neveola, zeng, olivier}@lhc.nlm.nih.gov

***Objective**: This paper explores alternative approaches for the evaluation of an automatic indexing tool for MEDLINE, complementing the traditional precision and recall method. **Materials and methods**: The performance of MTI, the Medical Text Indexer used at NLM to produce MeSH recommendations for biomedical journal articles is evaluated on a random set of MEDLINE citations. The evaluation examines semantic similarity at the term level (indexing terms). In addition, the documents retrieved by queries resulting from MTI index terms for a given document are compared to the PubMed related citations for this document. **Results**: Semantic similarity scores between sets of index terms are higher than the corresponding Dice similarity scores. Overall, 75% of the original documents and 58% of the top ten related citations are retrieved by queries based on the automatic indexing. **Conclusions**: The alternative measures studied in this paper confirm previous findings and may be used to select particular documents from the test set for a more thorough analysis.*

INTRODUCTION

The increasing number of electronic documents published in health-related journals and conferences makes the use of automatic tools necessary to ensure that reference databases such as MEDLINE can be kept up to date. In fact, the National Library of Medicine anticipates the need to index over 1 million journal articles annually by 2015, which is almost twice the number of articles processed in 2004. To accommodate the significant additional workload, indexers expect to rely on efficient indexing recommendations from NLM's automatic MeSH indexing system Medical Text Indexer (MTI) [1]. A key step to improving MTI and its use by indexers consists in accurately assessing the performance of the system and identifying its strengths and weaknesses.

The problem with evaluating indexing is that there is no universal reference indexing [2]. A given document may be correctly represented by different sets of descriptors. Many studies evaluating automatic systems use manual indexing as a gold standard [3-4]. Although it can be assumed that a set of descriptors produced by a human expert constitutes high quality indexing for a document, this set will only account for one of the possible sets that correctly represent the document. In fact, the results of human indexing consistency studies [5] support the idea that using manual indexing as a fixed gold standard results in drastically underestimating other valid indexing solutions.

For this reason, we have decided to explore alternative evaluation measures that take into account the possible variety of valid indexing sets for a given document. In this context, we assume that two valid sets of descriptors should not necessarily be the same, but should be semantically related. Moreover, they should result in a similar semantic position of the document they describe within the more general framework of the document collection. This study on evaluation measures for indexing is designed in continuity with previous work [4] in that the automatic indexing performance is assessed both at the term and document retrieval level. Our contribution is to complement the use of precision and recall techniques by considering the gold standard as one valid indexing set as opposed to a universal reference.

BACKGROUND

The MeSH thesaurus

The MeSH thesaurus is the controlled vocabulary used to index factual information in the biomedical domain. It is specifically used to index documents included in the MEDLINE database. It contains over 23,000 descriptors (main headings) organized in sixteen hierarchical tree structures. Each tree contains up to eleven levels denoting aboutness relationships between the terms. Some terms are shared by several trees. For example, the term *Alcohol-Related Disorders* belongs to both the *Diseases* (C) and *Psychiatry and Psychology* (F) trees.

Semantic similarity between MeSH descriptors

In the context of this study, semantic similarity refers to the similarity between two nodes in a taxonomy. The taxonomy under investigation in this study is MeSH. Unlike traditional edge-counting techniques,

semantic similarity methods are based on the assumption that the more information two terms share in common, the more similar they are. *Lin's similarity model* [6], for example, has been shown to produce relevant results when applied to the Gene Ontology [7].

Given two terms m_i and m_j, the Lin similarity between them is defined as:

$$sim\ (m_i, m_j) = \frac{2 \times \max_{m \in S(m_i, m_j)} [\log(\ p(m))]}{\log(\ p(m_i)) + \log(\ p(m_j))} \qquad (1)$$

where $S(m_i,m_j)$ represents the set of ancestor terms shared by both m_i and m_j, 'max' represents the maximum operator, and $p(m)$ is the probability of finding m or any of its descendants in a reference corpus (here, the probability of finding m as an index term in the entire MEDLINE collection). It generates normalized similarity values between 0 and 1. Because MeSH has a polyhierarchical structure (i.e., a term may belong to several trees), it is first partitioned into its 16 disjoint trees (A, B, …, Z). Therefore, term-term similarity across trees is 0, because there are no common ancestors between disjoint trees.

Because Lin's similarity model relies on information content, when one term is the parent of another, their similarity is low when the parent term is high in the hierarchy. Conversely, it is high when the parent term is low in the hierarchy. Figure 1 illustrates this phenomenon.

Biological Markers [D23.101]
  Antigens, Differentiation [D23.101.100]
    Antigens, CD [D23.101.100.110]
      Antigens, CD3 [D23.101.100.110.103]

Sim(Antigens, CD3, Antigens, CD) = 0.76
Sim(Antigens, CD3, Biological Markers) = 0.68

Figure 1 – Semantic similarity between MeSH terms from the Biological Markers subtree

The NLM Medical Text Indexer (MTI)

MTI recommendations result from the combination of two MeSH Indexing methods. These methods are a Natural Language Processing approach based on MetaMap UMLS Indexing [7] and a statistical, knowledge-based approach [1]. A clustering algorithm then produces a single ranked list of recommended MeSH terms by combining the recommendations from both methods using term weights, co-occurrence information, and whether the term was found in the title or not. Post-processing of this final ranked list involves a set of rules designed to filter out irrelevant indexing recommendations. Three levels of filtering are possible, depending on the desired trade-off between precision and recall. MTI is currently used in the NLM's Web-based Data Creation and Maintenance System (DCMS) as a semi-automatic Indexing tool. When indexing a document, the NLM indexers can view MTI recommendations obtained with "medium" filtering (on average 15 main headings per document) and select the ones they wish to use for the citation from a clickable list.

MATERIALS AND METHODS

Materials

Selecting Random Samples from MEDLINE. Because of the backlog in MEDLINE indexing, some of the articles from 2005 might not have been indexed yet as of early March 2006. In order to obtain the representative samples, we collected three random samples of about 5710 citations (1%) among the 571,304 that were indexed for MEDLINE in 2004. All three samples were used to evaluate the term similarity. One sample was used to evaluate document retrieval. The current version of MEDLINE relies on the 2006 version of MeSH.

Obtaining MTI Output. MEDLINE citations were processed by MTI with medium filtering.

Overview of the evaluation method

At the *term level*, automatic indexing is assessed by comparing Dice Similarity and Semantic Similarity between the set of MTI indexing recommendations and the MEDLINE indexing gold standard. At the *document level*, automatic indexing is assessed by computing the overlap between the PubMed Related Citations and the citations retrieved by a query composed from MTI recommendations.

Evaluation at the term level: Similarity measures

In order to compare the sets of indexing terms from MEDLINE and MTI, we compute both traditional and semantic similarity.

For traditional similarity, we use the *Dice measure*:

$$sim_{\ Dice} = \frac{2 \times AB}{A + B} \qquad (2)$$

where A and B represent the cardinality of the two sets of indexing terms and AB represents the cardinality of their intersection.

For *semantic similarity*, we aggregate the semantic similarity values obtained between terms. In practice, for two sets of MeSH terms I_A and I_B comprising A and B terms respectively, the semantic similarity between sets is defined as follows:

$$SS(I_A, I_B) = \frac{1}{A+B} \times \left(\sum_i \max_j (sim(m_i, m_j)) + \sum_j \max_i (sim(m_i, m_j))\right) \qquad (3)$$

where $sim(m_i,m_j)$ are computed using (1). This metric can be understood as a variant of the Dice similarity coefficient adapted to semantic similarity, in which the presence of a term in the intersection is replaced by the highest value of semantic similarity for each term to any term in the other set.

A detailed example for semantic similarity and Dice similarity is shown below. Table 1 presents the indexing sets.

MTI	MEDLINE
• *Calcifediol* • *Chromatography, Liquid* • *Spectrum Analysis, Mass* • *Quinoxaline*	• *Calcifediol* • *Chromatography, High Pressure Liquid* • *Spectrometry, Mass, Electrospray Ionization* • *Humans* • *Reference Standards*

Table 1 – MTI and MEDLINE indexing sets for the citation with PMID 15277348

	MTI	MEDLINE	Dice	S. S.
(MTI, MEDLINE) pairs	*Calcifediol*	*Calcifediol*	1	1
	Chromatography, Liquid	*Chromatography, High Pressure Liquid*	0	0.86
	Spectrum Analysis, Mass	*Spectrometry, Mass, Electrospray Ionization*	0	0.82
	Quinoxalines	-	0	0
(MEDLINE, MTI) pairs	*Calcifediol*	*Calcifediol*	1	1
	Chromatography, Liquid	*Chromatography, High Pressure Liquid*	0	0.86
	Spectrum Analysis, Mass	*Spectrometry, Mass, Electrospray Ionization*	0	0.82
	-	*Humans*	0	0
	-	*Reference Standards*	0	0
All pairs			2	5.36

Table 2 – Maximum similarity for each pair of terms from MTI and MEDLINE indexing sets (Dice and semantic similarity)

The total similarity shown in the bottom row of table 2 (columns 4 and 5) corresponds to the numerator of the similarity metric for Dice and semantic similarity, respectively. In both cases, the denominator is the sum of the cardinality of the two sets (here: 9).

Because the two sets only have one term in common, their Dice similarity is low (0.22). In contrast, the two sets have distinct, yet semantically related terms (e.g., *Chromatography, Liquid* and *Chromatography, High Pressure Liquid*). Therefore, the aggregated semantic similarity between the two sets is higher (0.59).

Evaluation at the document level

At the document retrieval level, we compare the MEDLINE citations retrieved by two methods. First, we use the indexing produced by MTI for a given document *D* to compose a query. Second, we use as gold standard the set PubMed Related Citations (PRCs) retrieved for *D*. Considering the PRC document set as our reference, we evaluate the overlap between the two document sets. Unlike the sets of documents returned by regular PubMed queries, PRCs are ranked by relevance. Therefore, we take into account the rank of each PRC in our analysis.

PubMed Related Citations (PRC). The PRC algorithm [4] computes a list of ranked citations based on the title, abstract, and MeSH index set of a document. For each document, the top related citation retrieved is the document itself, followed by citations whose content is most likely to be very similar to that of the original document. The program ELink from the PubMed E-Utilities was used to retrieve the related citations.

Query	Rank
((*a* AND *b* AND *c* AND *d*)) AND ct	1
((*a* AND *b* AND *d*) OR (*b* AND *c* AND *d*) OR (*a* AND *b* AND *c*) OR (*a* AND *c* AND *d*)) AND *ct*	2
((*c* AND *d*) OR (*a* AND *c*) OR (*a* AND *d*) OR (*a* AND *b*) OR (*b* AND *d*) OR (*b* AND *c*)) AND *ct*	3
((*a*) OR (*b*) OR (*c*) OR (*d*)) AND *ct*	4

Figure 2 – Sample of queries composed from MTI indexing of a document

Querying Against PubMed. PubMed does not rank results by relevance, but sorts the citations by publication date instead (by default). In order to compensate for the lack of relevance ranking in PubMed, we composed queries combining the index terms (except check tags) in such a way that the documents sharing a large number of index terms with the query would

rank higher than those sharing fewer terms. Check tags are appended to each query. For example, Figure 2 shows the queries composed from MTI indexing terms *a*, *b*, *c*, *d* and check tag *ct*, along with the order in which the corresponding results will be considered for relevance purposes.

RESULTS

At the term level: Dice *vs.* semantic similarity

Table 3 presents the similarity measures between sets of MTI and MEDLINE indexing terms for the three samples studied. Note that, for some citations (about 140 per sample), MTI does not produce any MeSH recommendations[1]. Those citations were excluded from the evaluation.

Number of citations	Average Dice similarity	Average semantic similarity
5567	0.398	0.541
5581	0.403	0.545
5573	0.399	0.541

Table 3 – Average Dice and *semantic similarity*

At the document level: Document retrieval

Table 4 and 5 present the overlap between the Related Citations (PRCs) and the set of documents retrieved by the queries composed from MTI indexing. For a given citation, the number of PRCs returned is, on average, 347.

Overall, 32% of all related citations retrieved for the 4009 original citations are also retrieved by the queries composed from MTI indexing. By convention, the top related citation is always the original document itself. In 3035 cases (75%), the original document is retrieved by the queries composed from MTI indexing. Because the number of citations returned by some queries composed from MTI indexing can be extremely large, we also investigated among how many citations the original document was retrieved. As shown in Table 4, the original document is retrieved among a maximum of 100 citations in 46% of the cases.

Considering only the top 10 related citations returned, we have a total of 23,494 citations for the 4009 queries. Overall, 58% of these top related citations were retrieved by the queries composed from MTI indexing regardless of the rank. More precisely, as shown in Table 5, the top related citations are retrieved among a maximum of 100 citations in 27% of the cases.

Top N documents	Percentile
Top 10	26%
Top 25	33%
Top 50	40%
Top 100	46%
Top 500	61%
Top 1000	68%

Table 4 – Original documents retrieved by queries composed from MTI indexing

Top N documents	Percentile
Top 10	9.6%
Top 25	15.1%
Top 50	21.4%
Top 100	27.4%
Top 500	46.1%
Top 1000	55.4%

Table 5 – Top 10 related citations retrieved by queries composed from MTI indexing

DISCUSSION

Benefit of using semantic similarity

We can see from Table 3 that in the three samples of citations, the average semantic similarity is higher than the average Dice similarity. This indicates that, in general, the terms recommended by MTI are indeed semantically related to gold standard terms. The example presented in the Methods section shows that these terms are often close ancestors or descendants of terms actually selected in the gold standard. In this example, MTI recommended *Chromatography, Liquid* when the MEDLINE indexers selected *Chromatography, High Pressure Liquid*, which is a direct child of *Chromatography, Liquid* in the E MeSH tree. This finding is compatible with previous observations [9] indicating that term specificity is one of the weakest points of automatic indexing systems.

In this particular example, the original document was retrieved among the top 25 documents returned, more specifically at rank 12. In this case, both semantic similarity and document set overlap concur to indicate that MTI recommended a valid set of indexing terms. Interestingly, there is a significant difference between the Dice and Semantic similarities (0.22 *vs.* 0.57). In general, large differences between Dice and semantic similarity are observed on documents exhibiting a large number of indexing terms. Therefore, the relaxed evaluation methods proposed in this study seem to be suitable for analyzing complex indexing

[1] This is often due to the unusual or metaphoric wording of the title and/or the absence of an abstract. For example, MTI did not produce recommendations for a journal article entitled "Sorry, we're all out." for which no abstract was available (PMID: 15584207).

sets where the choice of indexing terms specificity is more at stake.

Document retrieval

Overall, 75% of the original documents were retrieved, regardless of the rank. Noticeably, almost half of them were retrieved around rank 100. This is due in part to the ranking of the documents retrieved. Within each subquery resulting from combining indexing terms (Figure 2), documents were ranked by PMID, which means that the most recent documents appear first. Because our original documents were selected from documents published in 2004, they are less likely to appear in the top ranks. This indicates that the overall proportion of original documents or related citations retrieved is a more reliable indicator of the performance of MTI. The major finding of this study is that queries composed from MTI indexing terms retrieve the original document in 75% of the cases and retrieve the top 10 related citations of this document in 58% of the cases. These results indicate that, in the general framework of the MEDLINE collection, MTI indexing sets position documents in a semantic space close to that of MEDLINE indexing.

Limitations of the study

The experiment we report on was conducted using three samples of 5713 randomly selected citations from MEDLINE citations published in 2004, which corresponds to 3% of 2004 citations. Although this set of documents is representative of the MEDLINE collection, we plan to expand the study to a larger portion of MEDLINE in order to confirm our results.

Generalization of the evaluation methods

Precision and recall evaluation measures are widely used (e.g., at TREC conferences) and apply to indexing and information retrieval in virtually any field. In contrast, the semantic similarity measure we use was tailored for MeSH and based on frequency information from MEDLINE. However, the underlying principle (i.e., the semantic similarity between terms is based on their lowest common ancestor in a given taxonomy and on term frequency information derived from a reference corpus [6]) is generic and has been applied to other taxonomies, including WordNet and the Gene Ontology. Analogously, the search for related documents can be performed using a generic tool such as Google's "similar pages" feature.

CONCLUSIONS

In this paper, we have introduced alternative evaluation measures for automatic MeSH indexing, namely semantic similarity and document retrieval overlap. Applied to the NLM's indexing tool, document retrieval shows that MTI recommendations tend to position the document in the adequate semantic space within the MEDLINE collection. Semantic similarity brings out the documents which require a more complex indexing, and for which the choice of term specificity is a critical issue.

Acknowledgement

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and by an appointment of A. Névéol to the NLM Research Participation Program sponsored by NLM and administered by the Oak Ridge Institute for Science and Education. The authors would like to thank Alan Aronson, James Mork and Cliff Gay from the MTI team for providing practical help and insight on MTI, David Wheeler for his advice on querying against PubMed, and Sandy Tao for her groundwork on using MeSH semantic similarity for MTI evaluation.

References

1. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004: 268-72.
2. Lancaster, F. W. Indexing and abstracting in theory and practice. University of Illinois: Champaign, IL, 1991.
3. Névéol A, Mork JG, Aronson AR, Darmoni, SJ. Evaluation of French and English MeSH Indexing Systems with a parallel corpus. Proc AMIA Symp 2005.
4. Kim W, Aronson AR, Wilbur WJ. Automatic MeSH term assignment and quality assessment. Proc AMIA Symp. 2001:319-23.
5. Funk ME, Reid CA, McGoogan LS. Indexing consistency in MEDLINE. Bull. Med. Libr. Assoc. 1983:2 (71): 176-183.
6. Lin, D. An information-theoretic definition of similarity. In Proc. Int. Conf. on Machine Learning 1998:296-304.
7. Lord, P., Stevens, R., Brass, A. and Goble, C. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 2003:19, 1275-1283.
8. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001:17-21.
9. Névéol, A., Mary, V., Gaudinat, A., Boyer, C., Rogozan, A., & Darmoni, S. J. (2005). A Benchmark Evaluation of the French MeSH Indexing Systems. Proc AIME 2005:251-255.