Comparing a Rule-Based Versus Statistical System for Automatic Categorization of MEDLINE Documents According to Biomedical Specialty

Susanne M. Humphrey,* Aurélie Névéol, and Allen Browne
U.S. National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894. E-mail: {neveola, browne}@nlm.nih.gov

Julien Gobeil
Medical Informatics Service, University and University Hospitals of Geneva, CH-1211 Geneva 14, Switzerland. E-mail: julien.gobeill@sim.hcuge.ch

Patrick Ruch
BiTeM Group, Information Science Department, University of Applied Science, Geneva, 7 Drize, 1227 Carouge, Switzerland. E-mail: Patrick.Ruch@unige.ch

Stéfan J. Darmoni
CISMeF Group, Rouen University Hospital & GCSIS, LITIS EA 4108, Institute of BioMedical Research, University of Rouen, 1 rue de Germont, 76031 Rouen Cedex, France. E-mail: Stefan.Darmoni@chu-rouen.fr

Automatic document categorization is an important research problem in Information Science and Natural Language Processing. Many applications, including Word Sense Disambiguation and Information Retrieval in large collections, can benefit from such categorization. This paper focuses on automatic categorization of documents from the biomedical literature into broad discipline-based categories. Two different systems are described and contrasted: CISMeF, which uses rules based on human indexing of the documents by the Medical Subject Headings (MeSH) controlled vocabulary in order to assign metaterms (MTs), and Journal Descriptor Indexing (JDI), based on human categorization of about 4,000 journals and statistical associations between journal descriptors (JDs) and textwords in the documents. We evaluate and compare the performance of these systems against a gold standard of humanly assigned categories for 100 MEDLINE documents, using six measures selected from trec_eval. The results show that for five of the measures performance is comparable, and for one measure JDI is superior. We conclude that these results favor JDI, given the significantly greater intellectual overhead involved in human indexing and maintaining a rule base for mapping MeSH terms to MTs. We also note a JDI method that associates JDs with MeSH indexing rather than textwords, and it may be worthwhile to investigate whether this JDI method (statistical) and CISMeF (rule-based) might be combined and then evaluated showing they are complementary to one another.

Received February 12, 2009; revised May 15, 2009; accepted June 3, 2009

*Retired from U.S. National Library of Medicine. Current address: 2123 Arcola Avenue, Wheaton, MD 20902, USA. E-mail: susannehumphrey@yahoo.com

 DOI: 10.1002/asi.21170

Introduction

Categorization in the Biomedical Domain

This paper reports on a comparative evaluation of two methods of text categorization in the biomedical domain, where the categorization task consists of labeling documents according to biomedical specialty or discipline (e.g., Biochemistry, Cardiology, Epidemiology). Several other categorization tasks have been reported in the biomedical literature, including categorization into Medical Subject Headings (MeSH), Gene Ontology (GO), ICD-9, or SNOMeD categories (Aronson et al., 2007; Ehrler, Geissbühler, Jimeno, & Ruch, 2005; Ruch, 2006; Ruch, Gobeill, Lovis, & Geissbühler, 2008), but these tasks differ from categorization into biomedical specialties due to the nature of the categories used. That is, controlled vocabularies such as GO or ICD-9 include several thousand very specific "categories." Even where only a small subset of the categories is considered (e.g., Aronson et al., 2007), the degree of specificity of the categories makes the task very different from categorization into broad specialties. For example, the ICD-9 code "hematuria" is much more specific than the

corresponding biomedical specialty "urology." The fact that there is a limited number of biomedical specialties (about 100, as discussed in the next section) seems to favor using the broad range of machine-learning methods available for text categorization. However, as can be seen from an extensive review of these methods (Sebastiani, 2002) they require large sets of prelabeled documents for training. Such datasets are not available for biomedical specialties. Furthermore, as described in the evaluation section, creating gold-standard annotations is highly time-consuming and requires domain experts.

These issues were also discussed in previous reviews of categorization methods in the biomedical domain (Humphrey, 1999; Névéol et al., 2004).

Specific Categorization Methods in this Study

Two different systems that perform such categorization automatically were developed independently in the United States and in France. The Journal Descriptor Indexing (JDI) system, developed at the National Library of Medicine (NLM), categorizes text according to Journal Descriptor (JD) (Humphrey, 1998, 1999; Humphrey, Rogers, Kilicoglu, Demner-Fushman, & Rindflesch, 2006; Humphrey, Lu, Rogers, & Browne, 2006; National Library of Medicine, 2008a, 2008b). The Catalog and Index of Online Health Resources in French (CISMeF) system, developed at Rouen University Hospital, categorizes text according to Metaterm (MT) (CHU Hôpitaux de Rouen, 2008a, 2008b; Névéol, et al., 2004). JDs are a subset of subject terms from NLM's Medical Subject Headings (MeSH) used for describing journals per se (National Library of Medicine, 2008c) in NLM's List of Serials Indexed for Online Users (LSIOU) (National Library of Medicine, 2008d). MTs are terms for medical specialties or biological sciences selected by the CISMeF chief librarian (Douyère, et al., 2004).

Because of entirely different, separately developed approaches for categorizing text—manually maintained rules based on human indexing (for MT) versus statistical associations based on words in text (for JD), as described and illustrated further on—we thought it would be of interest to compare these two approaches by evaluating their performance for categorizing a set of 100 MEDLINE documents for which a human consensus of gold standard categorization was established. In both approaches, MTs/JDs are not assigned to text directly by indexers. Instead, MT categorization depends on MeSH terms assigned by indexers to the text to be categorized, and JD categorization depend on JDs assigned by a single indexer to about 4,100 journals per se (i.e., not the documents in the journals) in a serials database, a relatively modest, essentially one-time effort. A more detailed description of MT/JD categorization of text appears at the end of this section; in section "Categorization of a Sample MEDLINE document," MT/JD categorization is illustrated.

The list of 122 JDs (e.g., Biochemistry, Cardiology, Communicable Diseases, Complementary Therapies, Diagnostic Imaging, Environmental Health, Microbiology, Nursing, Public Health) is available on the Web (National Library of Medicine, 2008c). The list of 97 MTs in English (e.g., acupuncture, biochemistry, cardiology, diagnostic imaging, environment and public health, infectious diseases, microbiology, mycology, nursing care) is available on the Web (CHU Hôpitaux de Rouen, 2008c). Although there is considerable correspondence between MTs and JDs themselves, the respective methodologies and applications of MT and JD categorization are quite different.

MTs were designed for cataloging health resources available on the Internet; these resources have been cataloged by the CISMeF team using MeSH terms. In the current study, MT categorization is applied to MEDLINE documents based on MeSH terms assigned to them by NLM indexers. MT categorization has manually maintained rules that map MeSH terms—main headings (MHs) and subheadings (SHs)—in documents to MTs. For example, if a document is indexed with the MH *Heart Valve Diseases (the star means that this is a central concept in the document), it is automatically categorized under the MT Cardiology, because there is a rule that maps MeSH terms from cardiology hierarchies in MeSH (such as the Heart Diseases hierarchy, which contains Heart Valve Diseases) to the MT Cardiology. Points are assigned to the MT depending on whether or not the MH has a star. If the MH has a star, 100 points are assigned to the MT; otherwise 1 point is assigned.

On the other hand, JD categorization can be based on words in titles and abstracts. As will be explained by example in "Categorization of a Sample MEDLINE Document," JD categorization uses a dataset of 3 years of MEDLINE documents (the record for the document in NLM's PubMed database). JDs are not assigned to the documents in this dataset directly; each document in the dataset inherits (or imports) JDs from the journal in which it appears. As mentioned earlier, these JDs are manually assigned to journals in NLM's serials database. For example, all MEDLINE documents from the American Journal of Cardiology inherit the JD Cardiology from the record of this serial. In other words, a document from this journal is indexed under the JD Cardiology by virtue of the assignment of the JD Cardiology to this journal in the serials database. As a result, statistical associations can be made in the dataset between the words in MEDLINE documents in this journal and the JD Cardiology. These associations are then used for indexing documents outside the dataset; for example, a document in the *New England Journal of Medicine* containing many words associated with the *American Journal of Cardiology* in the dataset will be indexed automatically under the JD Cardiology.

Both systems are available on the Web (CHU Hôpitaux de Rouen, 2008d; National Library of Medicine, 2008b).

Benefits of Broad Categorization in the Biomedical Domain

There are many applications that can benefit from categorization into biomedical specialties, such as:

- Retrieval of resources in an online catalog (Gehanno, Thirion, & Darmoni, 2007), which was the original intent behind the development of MTs (Thirion & Darmoni, 1999).

DOI: 10.1002/asi

- WEB browsing by broad category. CISMef MT categorization is the precedent for the JAMA & ARCHIVES topic collections feature (American Medical Association, 2008; McGregor, 2005).
- Initial step in natural language processing (NLP). NLM's JD indexing (JDI) is used for identifying MEDLINE document in the molecular genetics domain before NLP begins (Névéol, Shooshan, Humphrey, Rindflesch, & Aronson, 2007; Rindflesch, Libbus, Hristovski, Aronson, & Kilicoglu, 2003).
- Initial step in gene symbol disambiguation (GSD). JDI can be used for identifying MEDLINE documents in the genetics domain (Hristovski, Peterlin, Mitchell, & Humphrey, 2005).
- Word sense disambiguation (WSD). JDI is the basis for semantic type indexing (STI) used for WSD, which has been described in detail (Humphrey, Rogers, Kilicoglu, Demner-Fushman, & Rindflesch, 2006), and is being investigated for MetaMap, a component of NLM's Medical Text Indexer (Aronson, Mork, Gay, Humphrey, & Rogers, 2004), formerly known as the Indexing Initiative System (Aronson, et al., 2000), which is in daily use to assist indexers in their indexing of documents for MEDLINE (Aronson, Mork, Lang, Rogers, & Névéol, 2008).
- A JDI-based method for automatic MeSH subheading attachment to main heading recommendations in NLM's Medical Text Indexer (Névéol, Shooshan, Humphrey, Mork, & Aronson, in press).
- Identifying the subdomains of a corpus for evaluation of that corpus. The corpus may be resources belonging to an institution (Darmoni et al., 2006) or problem lists detected in electronic medical records from an institution.

Categorization of a Sample MEDLINE Document Using MTs and JDs

To illustrate MT and JD categorization, we will use the following MEDLINE document, showing the PubMed Unique Identifier (PMID), title (TI), and MH (MeSH indexing terms):

PMID- 3181845
TI - Color Doppler echocardiography. Progress in the noninvasive diagnosis of heart valve diseases.
MH - Blood Flow Velocity
MH - Echocardiography, Doppler
MH - English Abstract
MH - Heart Defects, Congenital/diagnosis
MH - Heart Valve Diseases/*diagnosis
MH - Humans
MH - Image Interpretation, Computer-Assisted

MT Categorization

Table 1 shows the points assignments to MTs in the Metaterms column (Cardiology, etc.) from the MHs and the subheadings (SHs) in the MeSH terms column. An MH with a star (central concept) is known as a major MH; an SH with a star is known as a major SH. The scoring scheme is as follows:

- minor MH or SH: 1 point assigned for the mapped-to MT
- major MH or SH: 100 points assigned for the mapped-to MT

An MH with a starred SH is considered a major MH. MHs with more than one SH are counted twice. For example, MH1/*SH1/SH2 would be decomposed into two MH/SH pairs: MH1/*SH1 and MH1/SH2. Assume that indexing of a document is as follows:

MH1/*SH1/SH2
MH2/SH2

and that MH1, SH1, SH2, and MH2 map to metaterms MT1, MT2, MT3, and MT4, respectively.

The score for MT1 would be 101:

100 points for MH1 (with *SH1)
1 point for MH1 (with SH2)

The score for MT2 would be 100:

100 points for *SH1

The score for MT3 would be 2 points:

1 point for SH2 (appended to MH1)
1 point for SH2 (appended to MH2)

The score for MT4 would be 1 point:

1 point for MH2

If by some chance MH1, SH1, SH2, and MH2 all mapped to MT1, then the score for MT1 would be the sum of all the points, i.e., 204 points.

The MT categorization for the sample document is shown in the Metaterms and Final Metaterm score columns in Table 1. Because diagnosis and information science have no corresponding JD, these two MTs would be removed prior to performing the evaluation, resulting in the following MT categorization:

cardiology 103
diagnostic imaging 2
radiology 2
medical informatics 1
pediatrics 1
physiology 1

JD Categorization

To illustrate JD categorization we will use the words in the title (omitting stopwords): doppler, echocardiography, heart, noninvasive, and valve. For clarity, the calculations in this section are simplified; for example, it is essential that the JDI methodology use normalization techniques, rather than raw frequencies of words, in particular, a modified version of signal weight (Salton & McGill, 1983). Additional normalization is performed to counteract the effect of the uneven distribution of JDs in the training set (Humphrey, 1999). There are two type of categorization: based on word count and based on document count in the 3-year dataset. For example, to calculate the word count based score for the JD Cardiology for the word doppler, the system divides:

number of times doppler co-occurs with Cardiology in the dataset/number of times doppler occurs in the dataset

DOI: 10.1002/asi

TABLE 1. Points assignments to MTs according to MeSH indexing of document titled "Color Doppler echocardiography. Progress in the noninvasive diagnosis of heart valve diseases."

Metaterm	MeSH descriptor	Type of MeSH descriptor	MeSH score contribution	Final Metaterm score
cardiology	Blood Flow Velocity	MH	1	103
	Echocardiography, Doppler	MH	1	
	Heart Defects, Congenital	MH	1	
	Heart Valve Diseases	MH/*SH	100	
diagnosis	diagnosis	SH+SH*	101	104
	Blood Flow Velocity	MH	1	
	Echocardiography, Doppler	MH	1	
	Image Interpretation, Computer-Assisted	MH	1	
diagnostic imaging	Echocardiography, Doppler	MH	1	2
	Image Interpretation, Computer-Assisted	MH	1	
radiology	Echocardiography, Doppler	MH	1	2
	Image Interpretation, Computer-Assisted	MH	1	
medical informatics	Image Interpretation, Computer-Assisted	MH	1	1
pediatrics	Heart Defects, Congenital	MH	1	1
physiology	Blood Flow Velocity	MH	1	1
information science	Image Interpretation, Computer-Assisted	MH	1	1

To calculate the word count based score for the JD Diagnostic Imaging for the word doppler, the system divides:

number of times doppler co-occurs with Diagnostic Imaging in the dataset/number of times doppler occurs in the dataset

To calculate the document count based score for the JD Cardiology for the word doppler, the system divides:

number of documents in which doppler co-occurs with Cardiology in the dataset/number of documents in which doppler occurs in the dataset

To calculate the document count based score for the JD Diagnostic Imaging for the word doppler, the system divides:

number of documents in which doppler co-occurs with Diagnostic Imaging in the dataset/number of documents in which doppler occurs in the dataset

Of course, all word-JD scores from the dataset are precomputed, and are known as word-JD vectors.

Table 2 shows the JD scores for Cardiology and Diagnostic Imaging for words in the title of our sample document, and the JD categorization of the sample document, which is the average of the scores for each JD across the words. Table 3 shows JD categorization for the sample document. Note that the scores for Cardiology and Diagnostic Imaging are the average scores for the words according to Table 2.

This methodology can be described in terms of vectors. The 3-year dataset contains word-JD vectors for the words in the document in the dataset, where the JD vector for a word consists of the JD scores for that word, ordered alphabetically by JD. Knowing the JD scores for individual words, a document-JD vector of some document outside the dataset is the centroid of the JD vectors of the words in this document (i.e., the average of the scores across the words in the document). When we rank the JDs in this document-JD vector by score, we have the JD categorization of the document.

As with MT categorization, JDs with no corresponding MT were removed from the result. There are actually 122 JDs, but only 101 have corresponding MTs. Therefore, the JD categorization system was specially programmed for the evaluation not to return the 21 JDs having no corresponding MT.

Evaluation

Establishing a Gold Standard

In order to compare MT and JD categorization, a consensus of gold standard MTs/JDs was arrived at by two human experts (S.M. Humphrey and S.J. Darmoni[1]) for 100 documents that had been randomly selected from a month of MEDLINE documents indexed in January 1998 (for another project). We refer to these documents as our corpus.

Because there was no exact correspondence between the set of MTs and JDs, separate MT and JD consensus sets were compiled. Given that some MTs have no corresponding JD, and vice versa, these were eliminated from the set of MTs/JDs available for the consensus sets. In most cases, there was either exact (e.g., parasitology vs. Parasitology) correspondence or direct correspondence (oncology vs. Neoplasms), but allowances were made for near correspondence. For example, hepatology is an MT but not a JD, but gastroenterology/Gastroenterology is an MT/JD. Therefore, the MT consensus for a document in the field of hepatology

[1]Humphrey and Darmoni, experts in the JD and MT approaches, respectively, spent several weeks working to achieve the consensus by telephone and email. Their work was based on the title and abstract of the documents. About 2 weeks total were necessary to establish a correspondence between JDs and MTs. An additional 2 weeks were spent developing the gold standard consensus: both experts spent about 1 hour on each document. The documents were not categorized by either of their systems prior to their work, so that the gold standard was obtained independently from the automatic methods.

DOI: 10.1002/asi

TABLE 2. Scores for top two document JDs Cardiology and Diagnostic Imaging for words in document titled "Color Doppler echocardiography. Progress in the noninvasive diagnosis of heart valve diseases," and average scores across words, which are the scores for these JDs for the document.

Word and average across words	Word count-based method scores for top two JDs		Document count-based method scores for top two JDs	
	Cardiology	Diagnostic Imaging	Cardiology	Diagnostic Imaging
Doppler	0.029448	0.082110	0.066766	0.128493
Echocardiography	0.071619	0.047001	0.169341	0.095401
Heart	0.046655	0.005601	0.093659	0.014004
Noninvasive	0.016036	0.016944	0.046434	0.555228
Valve	0.107883	0.015819	0.153553	0.032634
Average score for JD	0.054328	0.033496	0.105951	0.065152

TABLE 3. JD categorization for document titled "Color Doppler echocardiography. Progress in the noninvasive diagnosis of heart valve diseases."

JD categorization based on word count			JD categorization based on document count		
Rank	Score	JD	Rank	Score	JD
1	0.054328	Cardiology	1	0.105951	Cardiology
2	0.033496	Diagnostic imaging	2	0.065152	Diagnostic imaging
3	0.032495	Pulmonary disease (specialty)	3	0.058277	Pulmonary disease (specialty)
4	0.026378	Vascular diseases	4	0.056590	Vascular diseases
5	0.016646	Surgery	5	0.030382	Surgery

was hepatology, and the JD consensus for the same document was Gastroenterology. If the MT system categorization was hepatology, this was counted as agreeing with the consensus (but not if it was gastroenterology), and if the JD system categorization was Gastroenterology, this was also counted as agreeing with the consensus. An example where the JD was more specific is the JD Drug Therapy and the corresponding MT therapeutics. If the JD consensus for a document was Drug Therapy, the MT consensus was therapeutics. Thus, if the JD system categorization was Drug Therapy, this was counted as agreeing with the consensus (but not if it was Therapeutics), and if the MT system categorization was therapeutics, this was also counted as agreeing with the consensus.

Evaluation Measures

The 100 documents were run through the respective MT and JD categorization systems and the trec_eval package was used for comparing the results (National Institute of Standards and Technology, 2008a). trec_eval was selected because it is well recognized in the Information Retrieval community, being the package used in Text Retrieval Conference (TREC) (National Institute of Standards and Technology, 2008b), and served well for our text categorization evaluation. In particular, being an off-the-shelf package, it obviated the need to develop programs to calculate and average the various precision and recall metrics we used. However, whereas trec_eval normally evaluates retrieval of documents relevant for topics, we used trec_eval to evaluate assignment of MTs/JDs for categorizing documents.

Many of the measures defined in the trec_eval package are defined and illustrated by Manning & Schütze (1999).

TABLE 4. MT and JD consensus for document titled "Association between p53 mutation and clinicopathological features of non-small cell lung cancer."

Consensus MTs for sample document	Consensus JDs for sample document
genetics	Genetics, medical
oncology	Neoplasms
pathology	Pathology
pulmonary disease (specialty)	Pulmonary disease (specialty)

In general, precision is the percentage of assigned MTs/JDs that are correct, i.e., of the MTs/JDs assigned to the document, what percentage is correct (matches the consensus). Recall is the percentage of correctly assigned MTs/JDs, i.e., of all the correct MTs/JDs for the document (in the consensus), what percentage has been assigned. Specifically, the following trec_eval metrics were selected, with definitions adapted to our categorization evaluation:

- R-prec (precision at the number of correct MTs/JDs). Precision at the position of the number of correct MTs/JDs (in the consensus).
- ircl_prn.0.00, or interpolated average precision at 0% (referred to as top precision in the remainder of this paper). The maximum of all precision measurements determined at each correct JD/MT assignment.
- P5 (precision at 5). Precision at five MTs/JDs assigned.
- P10 (precision at 10). Precision at ten MTs/JDs assigned.
- recall5 (recall at 5). Recall at five MTs/JDs assigned.
- recall10 (recall at 10). Recall at ten MTs/JDs assigned.

To illustrate, we use a sample document from our corpus titled "Association between p53 mutation and clinicopathological features of non-small cell lung cancer." Table 4 shows the MT and JD consensus for this document. Table 5 shows

DOI: 10.1002/asi

TABLE 5. MT and JD categorization results along with scores for document titled "Association between p53 mutation and clinicopathological features of non-small cell lung cancer." In the Cons. column, x denotes agreement with the consensus.

MTs for sample document			JDs (top 15 of 101) for sample document		
Cons.	Scores	MTs assigned	Cons.	Scores	JDs assigned
	621	anatomy	x	0.5863	Neoplasms
	418	physiology	x	0.5561	Pathology
x	415	genetics	x	0.3706	Genetics, Medical
x	314	oncology		0.2315	Molecular Biology
	208	cytology		0.2128	Cytology
	208	histology	x	0.2054	Pulmonary Disease (Specialty)
x	206	pathology		0.1746	Genetics
x	202	pulmonary disease (specialty)		0.1741	Gynecology
	16	statistics		0.1693	Urology
	9	epidemiology		0.1526	Gastroenterology
	4	environment and public health		0.1230	Surgery
	2	geriatrics		0.1217	Virology
				0.1199	Hematology
				0.1197	Radiology
				0.1161	Biochemistry

TABLE 6. Computation of trec_eval measures of MT and JD categorization results for document titled "Association between p53 mutation and clinicopathological features of non-small cell lung cancer." Figures in bold are results and are summarized in Table 7.

Measure	Computation and results for MTs for sample document	Computation and results for JDs for sample document
R-prec	2 (number of MTs assigned correctly at the 4th MT oncology) / 4 (number of MTs in the consensus) = **0.5000**	3 (number of JDs assigned correctly at the 4th JD Molecular Biology) / 4 (number of JDs in the consensus) = **0.7500**
Top precision	Precision at genetics = 1/3 = 0.3333 Precision at oncology = 2/4 = 0.5000 Precision at pathology = 3/7 = 0.4287 Precision at pulmonary disease (specialty) = 4/8 = 0.5000 Maximum = **0.5000**	Precision at Neoplasms = 1/1 = 1.0000 Precision at Pathology = 2/2 = 1.0000 Precision at Genetics, Medical = 3/3 = 1.0000 precision at Pulmonary Disease (Specialty) = 4/6 = 0.6667 maximum = **1.0000**
P5	Percentage MTs correct at 5th MT cytology = 2/5 = **0.4000** (maximum p5 = 4/5 = 0.8000)	Percentage JDs correct at 5th JD Cytology = 3/5 = *0.6000* (maximum p5 = 4/5 = 0.8000)
P10	Percentage MTs correct at 10th MT epidemiology = 4/10 = **0.4000** (maximum p10 = 4/10 = 0.4000)	Percentage JDs correct at 10th JD Gastroenterology = 4/10 = **0.4000** (maximum p10 = 4/10 = 0.4000)
recall5	Percentage of correct MTs at 5th MT cytology = 2/4 = **0.5000**	Percentage of correct JDs at 5th JD Cytology = 3/4 = *0.7500*
recall10	Percentage of correct MTs at 10th epidemiology = 4/4 = **1.0000**	Percentage of correct JDs at 10th JD Gastroenterology = 4/4 = **1.0000**

the MT and JD categorization results along with the scores for this document. Table 6 explains the computation of results of trec_eval measures for this document. Table 7 summarizes the results for this document.

To obtain results for the entire corpus for a particular MT or JD method, we average the respective measures across the documents in the corpus. For example, for MT categorization of the corpus:

document #1	
R-prec	0.0000
top precision	0.4000
P5	0.4000
P10	0.2000
recall5	0.6667
recall10	0.6667
document #2	
R-prec	1.0000
top precision	1.0000
P5	0.4000
P10	0.2000
recall5	1.0000
recall10	1.0000
...	
document #100	
R-prec	1.0000
top precision	1.0000

DOI: 10.1002/asi

TABLE 7. trec_eval measures of MT and JD categorization results for document titled "Association between p53 mutation and clinicopathological features of non-small cell lung cancer." Computation of these results is explained in Table 6.

Trec eval measures	Results for MTs for sample document	Results for JDs for sample document
R-prec	0.5000	0.7500
Top precision	0.5000	1.0000
P5	0.4000	0.6000
P10	0.4000	0.4000
recall5	0.5000	0.7500
recall10	1.0000	1.0000

P5 0.2000
P10 0.1000
recall5 1.0000
recall10 1.0000

average of 100 documents
R-prec 0.5577
top precision 0.8291
P5 0.4320
P10 0.2630
recall5 0.7127
recall10 0.8465

The results were further evaluated for statistical significance by the pairwise Wilcoxon test.

Results

Two versions of MT categorization and five versions of JD categorization were evaluated against the corresponding consensus, as shown in Table 8. These include the MT categorization, word count-based JD categorization, and document count-based categorization described and illustrated earlier.

The other MT categorization is known as MT majeurs, where "majeurs" refers to inclusion of only those MTs that are derived from starred (or major) MHs/SHs. Using the sample MEDLINE document for illustrating MT categorization earlier, the MT majeurs result would be as follows:

diagnosis 104
cardiology 103

and removing diagnosis, which has no corresponding JD, the result for our study would be:

cardiology 103

The three additional JD categorization methods use the fact that JD categorization can also use MHs/SHs in the MEDLINE document. The MH method uses only starred MHs/SHs in the document, and their statistical association with JDs. The Text MH WC method combines both words in titles/abstracts, employing the word count-based method, and starred MHs/SHs. The Text MH DC method combines both words in titles/abstracts, employing the document count-based method, and starred MHs/SHs.

We include these methods in our results for completeness, but our emphasis is on the MT method, which performs better than the MT majeurs method for all measures, and on the Text DC and Text WC methods because our main objective is to compare methodologies that require MeSH indexing (the MT method) against JD methodologies that do not use MeSH Indexing (JD Text WC and JD Text DC).

An exception, as discussed below in Future Work, might be comparing MT categorization to the JD categorization MH method.

We noted that in our evaluation of document categorization it is seldom the case that the highest attainable P5 and P10 is 1.0000. For example, P5 for document #2, as shown above, is 0.4000, which does not reflect the fact that this is the best possible P5, given that the consensus for this document consists of two MTs. It is not possible for P5 to be greater than $2/5 = 0.4000$. To give a perspective of P5, P10, and recall5 in relation to the highest attainable measures, we submitted to the trec_eval program a run for MT and a run for JD, where the MT and JD assignments, respectively, perfectly match the consensus. The results are in the MT highest attainable and JD highest attainable rows in Table 3. For those measures where the highest attainable score is not 1.0000, we follow the score by its percent of the highest attainable score. For example, the highest attainable score for P5 for the MT run that perfectly matches the consensus is 0.6300. The actual average P5 for the MT method is 0.4320, followed by (69%), which means that the P5 of 0.4320 is 69% of the highest attainable score of 0.6300.

TABLE 8. Average trec_eval measures across 100 documents for two MT methods and five JD methods. The percentage following a P5, P10, and recall 5 score indicates the percentage of highest attainable score, as discussed and illustrated in the Results section.

Method	R-prec	Top prec	P5	P10	recall 5	recall 10
MT	0.5577	0.8291	0.4320 (69%)	0.2630 (83%)	0.7127 (71%)	0.8465
MT majeurs	0.5468	0.8079	0.4140 (66%)	0.2270 (72%)	0.6777 (68%)	0.7310
MT highest attainable	1.0000	1.0000	0.6300	0.3160	0.9983	1.0000
JD Text WC	0.6077	0.9220	0.4500 (69%)	0.2740 (83%)	0.7127 (72%)	0.8418
JD Text DC	0.6167	0.9343	0.4520 (69%)	0.2750 (83%)	0.7125 (72%)	0.8422
JD MH	0.6135	0.9378	0.4600 (70%)	0.2860 (86%)	0.7310 (73%	0.8802
JD Text WC+MH	0.6468	0.9612	0.4680 (71%)	0.2840 (86%)	0.7427 (75%)	0.8703
JD Text DC+MH	0.6562	0.9495	0.4740 (72%)	0.2840 (86%)	0.7470 (75%)	0.8690
JD highest attainable	1.0000	1.0000	0.6560	0.3310	0.9950	1.0000

DOI: 10.1002/asi

Statistical Significance

We now present the results in terms of statistical significance of the results in Table 8 according to the Wilcoxon test, comparing the various MT and JD methods with one another for the six measures. Since the MT method is always superior to the MT majeurs method, we consider the MT method as representing the CISMeF system for comparison with the JD methods. Furthermore, we emphasize the JD Text WC and JD Text DC methods, since these methods do not require MeSH indexing, as does the MT method, but rather rely on words in titles and abstracts. Accordingly, the MT method, the JD Text WC, and JD Text DC methods are considered comparable for R-prec, P5, P10, recall5, and recall10. Only for top precision are these two JD methods superior to the MT method.

For completeness, we also compare the MT method with the other JD methods, which either rely entirely on MeSH indexing (JD MH) or combine reliance on MeSH indexing and words in titles and abstracts (JD WC+MH and JD DC+MH). For all measures, each of the methods that combine MeSH indexing and words in titles and abstracts are found to be superior to the MT method. The JD MH method is comparable to the MT method for R-prec and recall10, and superior to the MT method for top precision, P5, P10, recall5, and recall10.

We also note that, for all measures, JD methods that combine MeSH indexing and words in titles and abstracts are superior to the other JD methods based solely on words in titles and abstracts.

Discussion

Comparing Different Categorization Methods

There are obvious differences between the MT and JD categorization methods that impact the comparison of their performance. MT categorization produces only those MTs, with their scores, that the system deems appropriate for the document. JD categorization results in all JDs with their scores. We decided to impose a threshold of the top-scoring 15 JDs, which we feel accommodates the measures we selected from the trec_eval package. The top 15 is a practical threshold for applications using JD categorization.

The necessity to remove MTs with no corresponding JD, and vice versa, was mentioned earlier. Removal of MTs/JDs could affect performance. In fact, three documents had no results in the MT majeurs method because the MTs derived from starred MHs/SHs had no corresponding JD.

Tied scores are a problem for certain evaluation measures for the MT method. For example, the following is part of a result of the MT method for a document, where x denotes agreement of the MT with the consensus:

x physiology 907
x pharmacology 300
neurology 202
embryology 103
gynecology 101
x reproductive medicine 101
x obstetrics 101

Note the tied score of 101. According to trec_eval, P5 was 0.6000 (three correct of first five) for this document. However, as presented, disregarding score, only physiology and pharmacology are correct = 0.4000. trec_eval must have considered either reproductive medicine or obstetrics to be in the first five. According to trec_eval documentation, ties are broken deterministically; that is, regardless of the order of MTs with the same score for a given document, the P5 score will be the same.

The above problem areas are unavoidable due to the different nature of the MT and JD categorization methods. Nevertheless, we felt it was important to compare a categorization that depends on MeSH indexing with one that does not.

Little Manual Labor Required for Categorization Methods

Both the JD and MT categorization methods presented above are fully automatic. However, JD categorization relies on the one-time assignment of Journal Descriptors to each journal indexed for MEDLINE. The manual assignment of journal descriptors to MEDLINE journals is done independently from JD categorization. Similarly, MT categorization relies on the assignment of MeSH descriptors to documents. While it could be argued that indexing may be performed automatically using a tool such as MTI (Aronson et al., 2004) to reduce manual labor, as with JDI, manual MEDLINE indexing is performed independently from categorization. That is, in both cases, no manual labor is performed specifically for the categorization task. Rather, categorization uses the results of manual work performed anyway for other purposes.

The situation is slightly different where the semantic links between MeSH terms and MTs are concerned. While these links were originally developed for information retrieval (Névéol et al., 2004), their new application to categorization triggered a manual effort to optimize the network of links. Limited efforts are also devoted to updating and improving the links when new MeSH headings become available with new releases of thesaurus. However, this effort is estimated to be much smaller than the development of a large labeled corpus to be used for the training of machine learning methods.

Conclusions

We have presented a contrasted evaluation on MEDLINE documents of two methods of automatic categorization by biomedical specialty. We find that for most of the evaluation measures used in our study, the MT method (relying on MH manual indexing) and the JD method (relying on statistical processing of words in Title and Abstracts of documents) perform similarly. However, the JD method outperforms the MT

DOI: 10.1002/asi

method for one measure, top precision. These results favoring the JD method imply that not much is gained by MH indexing, especially taking into consideration the intellectual overhead of indexing, and maintaining the links between MT and MHs. We also find that JD methods that combine MHs and text words outperform JD methods relying on text words only. This is not surprising since MHs definitely add valuable semantic content to the document description.

Future Work

The results of this study can be used to improve the respective categorization methods—for instance, additional MH-MT mappings (e.g., MH Vena Cava Filters–MT cardiology) were created after this study by reviewing statistical associations between MHs and JDs. Another perspective would be to investigate whether some combination of the methods might result in better performance than either method alone. A particular opportunity to do this might be comparison of the MT method and the JD MH method, since both use only MeSH indexing. The question would be whether using statistical associations between starred MHs/SHs and JDs might be complementary to using MHs/SHs mapping rules to MTs. Because of the limited number of documents in the test corpus, it may be desirable to develop a larger corpus of documents with gold standard indexing to perform a further comparison of the MT and JD approaches. Another interesting study would be to use a consistency study corpus (Funk and Reid, 1983), where several equivalent MeSH indexing sets are available for each document in the corpus in order to study the robustness of the methods to indexing variation, without requiring a gold standard.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine; an appointment of A. Névéol to the Lister Hill Center Fellows Program; appointments of P. Ruch and S.J. Darmoni to the to the Lister Hill Center Visitors Program, sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

Reference**

American Medical Association (2008). JAMA & Archives Topic Collections. Retrieved July 1, 2009, from http://pubs.ama-assn.org/collections

Aronson, A.R., Bodenreider, O., Chang, H.F., Humphrey, S.M., Mork, J.G., Nelson, S.J., et al. (2000). The NLM Indexing Initiative. In Proceedings of the American Medical Informatics Association Annual Symposium. (pp. 17–21). Retrieved July 14, 2009, from http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=2243970&blobtype=pdf

Aronson, A.R., Bodenreider, O., Demner-Fushman, D., Fung, K.W., Lee, V.K., Mork, J.G., et al. (2007). From indexing the biomedical literature to coding clinical text: Experience with MTI and machine learning approaches. In Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing (BioNLP 2007) (pp. 105–112). Stroudsbourg, PA: Association for Computational Linguistics.

Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., & Rogers, W.J. (2004). The NLM Indexing Initiative's Medical Text Indexer. Studies in Health Technology and Informatics, 107(Pt 1), 268–272. Retrieved July 1, 2009, from http://skr.nlm.nih.gov/papers/references/aronson-medinfo04.wheader.pdf

Aronson, A.R., Mork, J.G., Lang, F.M., Rogers, W.J., & Névéol, A. (2008). NLM Medical Text Indexer: A tool for automatic and assisted indexing. NLM Technical Report No. LHNCBC-TR-2008-002. Bethesda, MD: U.S. National Library of Medicine, April 2008. Section 4.4 Word Sense Disambiguation. (pp. 12–13). Retrieved July 1, 2009, from http://lhncbc/lhc/docs/reports/2008/tr2008002.pdf

CHU Hôpitaux de Rouen (2008a). Catalogue et Index des Sites Medicaux Francophones. Retrieved July 1, 2009, from http://www.cismef.org

CHU Hôpitaux de Rouen (2008b). CISMeF: Catalog and index of French-language health internet resources. A quality-controlled subject gateway. Retrieved July 1, 2009, from http://www.chu-rouen.fr/cismef/cismefeng.html

CHU Hôpitaux de Rouen (2008c). Additional subject subset for PubMed. Retrieved July 1, 2009, from http://doccismef.chu-rouen.fr/liste_des_meta_termes_anglais.html

CHU Hôpitaux de Rouen (2008d). MeSH categorization. Retrieved July 1, 2009, from http://documvf.crihan.fr/servlets/Categoriseur

Darmoni, S.J., Névéol, A., Renard, J.M., Gehanno, J.F., Soualmia, L.F., Dahamna, B., et al. (2006). A MEDLINE categorization algorithm. BMC Medical Informatics and Decision Making, 6, 7. Retrieved July 14, 2009, from http://biomedcentral.com/1472-6947/6/7

Douyère, M., Soualmia, L.F., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J.P., et al. (2004). Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Information and Libraries Journal, 21(4), 253–261. Retrieved July 1, 2009, from http://www3.interscience.wiley.com/cgi-bin/fulltext/118813886/PDFSTART

Ehrler, F., Geissbühler, A., Jimeno, A., & Ruch, P. (2005). Data-poor categorization and passage retrieval for gene ontology annotation in Swiss-Prot. BMC Bioinformatics. 6 Suppl 1:S23. Retrieved July 14, 2009, from http://biomedcentral.com/1471-2105/6/S1/S23

Funk, ME., & Reid, CA. (1983). Indexing consistency in MEDLINE. Bulletin Medical Library Association, 71(2), 176–183.

Gehanno, J.F., Thirion, B., & Darmoni, S.J. (2007). Evaluation of meta-concepts for information retrieval in a quality-controlled health gateway. In Proceedings of the American Medical Informatics Association (pp. 269–273). Retrieved July 14, 2009, from http://telemedicina.unifesp.br/pub/AMIA/2007%20AMIA%20Proceedings/data/papers/papers/AMIA-0085-S2007.pdf

Hristovski, D., Peterlin, B., Mitchell, J.A., & Humphrey, S.M. (2005). Using literature-based discovery to identify disease candidate genes. International Journal of Medical Informatics, 74(2–4), 289–298.

Humphrey, S.M. (1998). A new approach for automatic indexing using journal descriptors. In C.M. Preston (Ed.), Proceedings of the 61st ASIS Annual Meeting (pp. 496–500). Medford, NJ: Information Today.

Humphrey, S.M. (1999). Automatic indexing of documents from journal descriptors: A preliminary investigation. Journal of the American Society for Information Science, 50(8), 661–674.

Humphrey, S.M., Rogers, W.J., Kilicoglu, H., Demner-Fushman, D., & Rindflesch, T.C. (2006). Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. Journal of the American Society for Information Science and Technology, 57(1), 96–113. Erratum in: (2006), 57(5), 726.

Humphrey, S.M., Lu, C.J., Rogers, W.J., & Browne, A.C. (2006). Journal Descriptor Indexing tool for categorizing text according to discipline or semantic type. In Proceedings of the American Medical Informatics

**Regarding references with PMCID: When a PMCID is searched in NLM's PubMed, the reference is retrieved with a link to the free full text article in PubMed Central.

DOI: 10.1002/asi

Association (p. 960). Retrieved July 14, 2009, from http://www. pubmedcentral. nih.gov/picrender.fcgi?artid=1839325&blobtype=pdf

Lu, C.J., Humphrey, S.M., & Browne, A.C. (2008). A method for verifying a vector-based text classification system. In Proceedings of the American Medical Informatics Association. Retrieved July 14, 2009, from http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/tc/current/docs/userDoc/references/Training%20Set%20Submit.pdf

Manning, C.D., & Schütze, H. (1999). Foundations of statistical natural language processing (pp. 268, 269, 534–538). Cambridge, MA: MIT Press.

Masseroli, M., Kilicoglu, H., Lang, F.M., & Rindflesch, T.C. (2006). Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. BMC Bioinformatics, 7, 291. Retrieved July 14, 2009, from http://biomedcentral.com/1471-2105/7/291

McGregor, B. (2005). Constructing a concise medical taxonomy. Journal of the Medical Library Association, 93(1), 121–123. Retrieved July 14, 2009, from http://biomedcentral.nih.gov/articlerender.fcgi?artid=545132

National Institute of Standards and Technology (2008a). trec_eval. Retrieved July 1, 2009, from http://trec.nist.gov/trec_eval

National Institute of Standards and Technology (2008b). Text Retrieval Conference (TREC). Retrieved July 1, 2009, from http://trec.nist.gov

National Library of Medicine (2008a). Journal Descriptor (JD) Indexing. Retrieved July 1, 2009, from http://ii/JDorig.shtml

National Library of Medicine (2008b). Text categorization. Retrieved July 1, 2009, from http://specialist.nlm.nih.gov/tc

National Library of Medicine (2008c). Journal subject terms. Retrieved July 1, 2009, from http://www.nlm.nih.gov/bsd/journals/subjects.html

National Library of Medicine (2008d). List of serials indexed for online users. Retrieved July 1, 2009, from http://www.nlm.nih.gov/tsd/serials/lsiou.html

Névéol, A., Soualmia, L.F., Douyère, M., Rogozan, A., Thirion, B., & Darmoni, S.J. (2004). Using CISMeF MeSH "encapsulated" terminology and a categorization algorithm for health resources. International Journal of Medical Informatics, 73(1), 57–64.

Névéol, A., Shooshan, S.E., Humphrey, S.M., Mork, J.G., & Aronson R. (In press). A recent advance in the automatic indexing of the biomedical literature. Journal of Biomedical Informatics.

Rindflesch, T.C., Libbus, B., Hristovski, D., Aronson, A.R., & Kilicoglu, H. (2003). Semantic relations asserting the etiology of genetic diseases. In Proceedings of the American Medical Informatics Association (pp. 554–558). Retrieved July 14, 2009, from http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1480275&blobtype=pdf

Ruch, P. (2006). Automatic assignment of biomedical categories: Toward a generic approach. Bioinformatics 22(6), 658–664.

Ruch, P., Gobeill, J., Lovis, C., & Geissbühler, A. (2008). Automatic medical encoding with SNOMED categories. BMC Medical Informatics and Decision Making, 27(8) Suppl 1:S6. Retrieved July 14, 2009, from http://biomedcentral.com/1472-6947/8/S1/S6

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval (pp. 63–66). New York: McGraw-Hill.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.

Thirion, B., & Darmoni, S.J. (1999). Simplified access to MeSH tree structures on CISMeF. Bulletin of the Medical Library Association, 87(4), 480–481. Retrieved July 14, 2009, from http://pubmedcentral.nih.gov/articlerender.fcgi?artid=226675

DOI: 10.1002/asi