Identifying Publication Types Using Machine Learning

Antonio J. Jimeno Yepes[1,2], James G. Mork[2], Alan R. Aronson[2]

[1]NICTA Victoria Research Lab, Melbourne, Australia
antonio.jimeno@gmail.com
[2]National Library of Medicine, Bethesda, MD, USA
{mork,alan}@nlm.nih.gov

Abstract. Every year the number of journals and the number of articles to be indexed grows at the U.S. National Library of Medicine (NLM) causing an ever increasing demand on the highly qualified, but, relatively small, dedicated staff of indexers. We present a methodology for identifying MeSH (Medical Subject Headings) Publication Types for assisting the indexers in the categorization of these MEDLINE citations. Publication Types are used by the indexer to describe the type or genre of an article instead of what the article is about, making this a different kind of text categorization problem from identifying MeSH Descriptors. Our goal is to apply a machine learning approach to recommending Publication Types which will save indexers time by providing a precise list of possible Publication Types for each article. Our experiments involved several different machine learning methods to provide Publication Type recommendations which were then evaluated against the gold standard of human indexing. Our results show that machine learning in most cases adds a great deal to the overall performance of recommending Publication Types. Our experiments also show that in some cases, either the full text of the article or feature engineering will be required to accurately produce some Publication Type recommendations.

Keywords: Indexing methods, Text categorization, Machine learning, MeSH, MEDLINE

1 Introduction

The MEDLINE®/PubMed® database contains over 21 million citations[1]. It currently grows at the rate of around 800,000 indexed citations per year covering almost 6,000 international biomedical journals[2] in 58 languages. These new citations are manually indexed by a relatively small, dedicated staff of indexers at the U.S. National Library of Medicine (NLM). In this paper, we will use the terms *article* and *citation* interchangeably, but they do refer to two distinct entities in the indexing world. Indexers index from the full text of an *article*, and the results of that effort along with the title and abstract from the *article* are stored as a *citation* in the MEDLINE/PubMed data-

[1] http://mbr.nlm.nih.gov
[2] www.nlm.nih.gov/bsd/bsd_key.html

base. The indexers use the Medical Subject Headings (MeSH®) [3]controlled vocabulary to summarize the central points of full text articles. The 2013 MeSH vocabulary consists of 26,853 MeSH Descriptors[4] which are further qualified by a set of 83 Mesh Qualifiers (Subheadings). For example, *Aspirin/therapeutic use* illustrates the MeSH Descriptor *Aspirin* being qualified by the MeSH Qualifier *therapeutic use* showing that the article is not about *Aspirin* in general, but, more specifically about the *therapeutic uses* of *Aspirin*. There are also 214,816 Supplementary Concepts available to the indexer for detailing important chemicals, drugs, or proteins identified in the articles. In addition to summarizing the main points of each article, the indexer is also responsible for other curation tasks such as assigning one or more Publication Types which define the genre of the article.

Publication Types (PTs)[5] are a special type of MeSH Heading that are used to indicate what an article is rather than what it is about. There are 61 PTs identified in the four MeSH Publication Characteristics (V) Tree top-level sub-trees that the indexers typically use. These four sub-trees describe a wide range of document types or genres for PTs: *Publication Components [V01]* (e.g., Architectural Drawings), *Publication Formats [V02]* (e.g., Eulogies), *Study Characteristics [V03]* (e.g., Clinical Trial), and *Support of Research [V04]* (U.S. Government and non-U.S. Government) with some PTs included in multiple sub-trees. Multiple PTs can be assigned to the same article by the indexer.

The ever increasing demand for indexing (502,056 indexed in 2002 to 760,903 indexed in 2012, and with NLM expecting to index over one million articles annually within a few years) is a growing and burdensome workload in a time of dwindling resources. NLM created the NLM Indexing Initiative (II) [1] project to explore indexing methodologies that could assist indexers by providing tools to increase their productivity while maintaining their high quality of indexing. The II project has previously shown that the right tools can help significantly reduce the amount of time required to manually index articles: MetaMap [2] identifying Unified Medical Language System (UMLS) ® concepts in biomedical text, the NLM Medical Text Indexer (MTI) [3] providing indexing recommendations and acting as a First Line Indexer for a select number of journals, and our previous success with machine learning providing recommendations for twelve of the most commonly used MeSH Check Tags [4,5] in MTI with an 80% success rate.

In 2004, testing MTI's PT recommendations showed that MTI was not very good at the task, as shown in Table 1. MTI has two main methods of summarizing what a citation is about: MetaMap Indexing (MMI) [2] and the PubMed Related Citations (PRC) [6] algorithm. MMI analyzes the citation identifying Unified Medical Language System (UMLS) concepts that best match the text of the citation. MTI then maps these UMLS concepts to the MeSH vocabulary using the Restrict-to-MeSH [7]

[3] http://www.nlm.nih.gov/pubs/factsheets/mesh.html
[4] http://www.nlm.nih.gov/mesh/intro_record_types.html
[5] http://www.nlm.nih.gov/mesh/pubtypes.html

mappings, which are based primarily on the semantic relationships of the UMLS concepts. The PRC algorithm is a modified k-NN algorithm which relies on document similarity to identify potentially relevant MeSH Descriptors. Both of the MTI methods are focused on summarizing the contents of the citation and not on analyzing the type of document being processed which accounts for MTI's poor performance with PTs. MTI performed so poorly on PTs that it was not used for 46 of the 61 PTs from the beginning, and we stopped recommending the remaining 15 PTs altogether on November 10, 2004.

Table 1 shows the MTI performance as of November 10, 2004 for the fifteen PTs that MTI was recommending at the time. This was a time of transition where some MeSH Descriptors were being designated as Publication Types, so the vast majority of these fifteen terms were actually MeSH Descriptors transitioning to Publication Types; they are denoted as MP in the Type column. Table 1 also shows the frequency of each term being used by the human indexer (Index), recommended by MTI (MTI), where the two matched (Match), Precision (P), Recall (R), and F_1 measure for each of the fifteen terms. Four of these fifteen terms (*Congresses*, *English Abstract*, *In Vitro*, and *Meta-Analysis*) are included in our current study and are highlighted in the table. *Congresses* performs the best of these four with an F_1 of 0.3397, while *English Abstract* has by far the best Precision (1.0000) with a correspondingly poor Recall (0.0005). Both *Meta-Analysis* and *In Vitro* have similar results with very high Recalls (0.8403 and 0.9969, respectively) and very low Precisions (0.1590 and 0.0917, respectively).

Table 1. Historical MTI baseline performance for current Publication Types

Term	Type	Index	MTI	Match	P	R	F_1
Bibliography	MP	48	1,009	24	0.0238	0.5000	0.0454
Biography	MP	920	37	18	0.4865	0.0196	0.0376
Congresses	MP	209	521	124	0.2380	0.5933	0.3397
Directory	PT	3	4	0	0.0000	0.0000	0.0000
Duplicate Publication	MP	1	99	1	0.0101	1.0000	0.0200
English Abstract	MP	1,954	1	1	1.0000	0.0005	0.0010
Evaluation Studies	MP	1,317	4,245	384	0.0905	0.2916	0.1381
Government Publications	MP	0	12	0	0.0000	0.0000	0.0000
In Vitro	MP	2,618	28,468	2,610	0.0917	0.9969	0.1679
Legislation	MP	9	561	7	0.0125	0.7778	0.0246
Meta-Analysis	MP	263	1,390	221	0.1590	0.8403	0.2674
Multicenter Study	PT	192	51	34	0.6667	0.1771	0.2798
Portraits	MP	495	1,553	443	0.2853	0.8949	0.4326
Retraction of Publication	PT	6	51	5	0.0980	0.8333	0.1754
Twin Study	PT	21	26	18	0.6923	0.8571	0.7660
Overall		8,056	38,028	3,890	0.1023	0.4829	0.1688

Our goal now is to consider the task of recommending PTs as a text categorization task using machine learning, which could save indexers even more time by providing a precise list of possible PTs for each article. There is no previous work on using machine learning in the context of PTs, though review of existing work for MeSH indexing [4,5,8,9] illustrates many cases where machine learning has been applied effectively. In addition, a large corpus of indexed MEDLINE citations is available as training data. There are several challenges to our approach:

1. The indexers index from the full text of an article in making their determinations of which PTs to assign while we are currently limited by license restrictions to just the title and abstract found in the MEDLINE citation.
2. Inconsistency between MeSH indexers [10] due to different interpretations of the article and different understanding of MeSH could result in an inconsistent gold standard and provide less than optimal training for the algorithms.
3. Changes to the indexing policy over time can introduce inconsistencies in the machine learning training. For example, if we have trained with years 2010, 2011, and 2012 and a new Publication Type was added in 2011, we have the potential for inconsistencies in the 2010 training data due to articles that look like they should have the new Publication Type assigned, but, do not. To help limit this problem, we have created a training set with MEDLINE citations from the last three years.
4. 18 of the 61 Publications Types commonly used by the indexers are found in multiple Publication Characteristics MeSH tree sub-trees. For example, the Publication Type *Letter* appears in the *Publication Components* (V01) and *Publication Formats* (V02) sub-trees. This presents a possible ambiguity problem and at the very least introduces possibly confusing documents for the machine learning training.

2 Methods

We have studied the use of various machine learning algorithms testing their ability to accurately recommend several different types of PTs for MEDLINE citations. We have selected the following ten PTs to see if we could provide reliable recommendations.

- **Case Reports**: Clinical presentations that eventually lead to a diagnosis.
- **Clinical Trial**: Work that is the report of a pre-planned clinical study.
- **Congresses**: Published records of the papers delivered at or issued on the occasion of individual congresses, symposia, and meetings.
- **Controlled Clinical Trial (CCT)**: Work consisting of a clinical trial involving one or more test treatments and at least one control treatment.
- **Editorial:** Work consisting of a statement of the opinions, beliefs, and policy of the editor or publisher of a journal.

- **English Abstract:** English Abstracts of foreign articles.
- **In Vitro:** Studies using excised tissues.
- **Meta-Analysis:** Work consisting of studies using a quantitative method of combining the results of independent studies.
- **Randomized Controlled Trial (RCT):** Similar to Controlled Clinical Trial, but requires that the treatments to be administered are selected by a random process.
- **Review:** An article or book published after examination of published material on a subject.

These ten PTs were selected because they represent some of the most frequently used PTs and provide a good cross category sample of the four Publication Characteristics MeSH tree sub-trees for PTs. We also limited our set to 10 PTs to facilitate training and evaluation.

As mentioned before, changes in the indexing policy can have a dramatic effect on how articles are indexed and can create inconsistencies in a large training corpus if special care is not taken. To reduce the chance of this, we have focused on the last three full indexing years using the 2012 MEDLINE Baseline. We used the Medline Baseline Repository Query Tool[6] to identify a list of PMIDs (PubMed Unique Identifiers) for Date Completed (date indexing was applied to the citation) ranging from January 1, 2009 to December 31, 2011. The Query Tool also allowed us to randomly divide the list of PMIDs into Training (2/3) and Testing (1/3) sets. We ended up with 1,784,061 randomly selected PMIDs for Training and 878,718 for Testing. Once we had the two lists of PMIDs, we extracted the actual citations from the 2012 MEDLINE Baseline in XML format for use with our MTI ML machine learning package[7]. The MTI ML package was developed as part of the Indexing Initiative effort to provide machine learning algorithms optimized for large text categorization tasks and capable of combining several text categorization solutions. It is available subject to the MetaMap Terms and Conditions[8].

Certain types of articles require special indexing. For example, a *Comment On* article, which is an article commenting on a different article, is indexed by simply using the indexing from the originating article. For a *Review* type of article, the indexer uses fewer MeSH Headings that tend to be more general in nature than they would use for a non-Review article. For these reasons, when we assembled the final data set, we also filtered out the articles requiring special handling to create as clean a data set as possible. Specifically, we removed the following types[9] of articles from our data sets: *OLDMEDLINE*, *PubMed-not-MEDLINE*, articles with no indexing, *CommentOn*, *RetractionOf*, *PartialRetractionOf*, *UpdateIn*, *RepublishedIn*, *ErratumFor*,

[6] http://mbr.nlm.nih.gov
[7] http://ii.nlm.nih.gov/MTI_ML/index.shtml
[8] http://metamap.nlm.nih.gov/MMTnCs.shtml
[9] http://www.nlm.nih.gov/bsd/licensee/elements_alphabetical.html

and *ReprintOf*. This left us with 1,321,512 articles for Training and 651,617 articles for Testing. The data sets used for these experiments are available from our Indexing Initiative Data Sets and Test Collections web page[10].

The task of assigning PTs to a MEDLINE citation can be seen as a text categorization task [4,8], in which the PTs are the categories to be assigned. In our experiments, we have trained binary classifiers to predict if the article should be indexed with a given PT or not. We have selected several learning methods in these experiments focusing on learning methods that can be trained in a reasonable time due to the large number of citations under consideration. Among these methods are a linear SVM implementation based on Hinge Loss and Huber Loss and an implementation of AdaBoostM1 that uses decision trees as base learner. In addition, we have considered Naïve Bayes and Logistic regression from the Mallet[11] package.

SVM has been shown to perform well on text categorization tasks [11]. We have used an implementation of SVM with linear kernel based on Hinge loss and stochastic gradient descent and modified Huber loss proposed by Zhang's [12] work used by Yeganova et al. [13], which has been shown to improve the performance of Hinge loss in the case of very imbalanced training sets. It is a wide margin classifier with a quadratic loss function. We have restricted our study to linear kernels due to the size of our data sets, but it would be worth exploring efficient implementations for learning with more complex kernels.

One of the algorithms that we have extensively used is AdaBoostM1 (Ada) using an implementation of decision trees based on C4.5 as the base learning algorithm. In previous work, Ada had performed well on the Check Tags set [8,9], and we were interested in evaluating its performance with a larger, more diverse set of terms. Our implementation of C4.5 relies on binary features, which provide a more efficient implementation of the decision tree in terms of memory and time required for training.

The SVM and AdaBoostM1 implementations are available from the MTI ML package[12], which has been used in several MeSH indexing research efforts and has become part of the MTI system. The MTI ML tool is already configured to work with MEDLINE citations and provides several configuration options to deal with different MEDLINE citation fields. The MTI ML package has also been extended to export the preprocessing of the articles for use by the Mallet package using its SVMLight[13] interface.

[10] http://ii.nlm.nih.gov/DataSets/index.shtml#2013_BioASQ
[11] http://mallet.cs.umass.edu
[12] http://ii.nlm.nih.gov/MTI_ML/index.shtml
[13] http://svmlight.joachims.org

3 Results

Table 2 depicts all of the PTs involved in our current study, the frequency of their true positive occurrences in the Test set (Occurs), the associated abbreviation (Abbrev) used in Table 3, and the Baseline F_1 where available from the earlier MTI results shown in Table 1.

Table 2. Baseline results and occurrences data for Publication Types test set and Table 3 Key

Publication Type	Occurs	Abbrev	Baseline F_1
Case Reports	51,037	CR	-
Clinical Trial	6,165	CT	-
Congresses	1,954	CO	0.3397
Controlled Clinical Trial	1,727	CC	-
Editorial	11,519	ED	-
English Abstract	46,471	EA	0.0010
In Vitro	4,284	IV	0.1679
Meta-Analysis	3,467	MA	0.2674
Randomized Controlled Trial	17,356	RC	-
Review	75,298	RV	-

Table 3. Publication Type Machine Learning F_1 Results by Method

Method	CR	CT	CO	CC	ED	EA	IV	MA	RC	RV
Mhl	0.7948	0.1204	0.6997	0.0578	0.1452	0.5770	0.1549	0.7093	0.7464	0.7324
Mhl-F	0.8131	0.1153	0.6999	0.0624	0.5426	0.8198	**0.1610**	0.7231	0.7544	0.7512
Mhl-B	0.8291	0.0993	**0.7113**	0.0192	0.2290	0.6386	0.1146	0.7687	0.7840	0.7485
Mhl-BF	0.8377	0.0909	0.7024	0.0192	**0.5584**	0.8318	0.1100	0.7733	0.7911	0.7660
Sgd	0.8075	0.0058	0.6918	0.0103	0.0844	0.5898	0.0734	0.7410	0.7732	0.7579
Sgd-F	0.8258	0.0943	0.7004	0.0380	0.3461	0.8255	0.1505	0.7310	0.7683	0.7685
Sgd-B	0.8252	0.0870	0.7109	0.0182	0.1183	0.6425	0.1049	**0.7742**	0.7899	0.7582
Sgd-BF	0.8392	0.0836	0.7089	0.0181	0.4939	0.8343	0.1005	0.7727	0.7910	0.7699
NB	0.6985	0.0281	0.4508	0.0009	0.0910	0.4215	0.1056	0.3125	0.4936	0.6355
NB-F	0.7461	0.0032	0.0652	0.0000	0.0889	0.5180	0.0012	0.0005	0.2544	0.5452
NB-B	0.7007	0.0000	0.0882	0.0000	0.0148	0.0857	0.0000	0.0000	0.0999	0.4330
NB-BF	0.6747	0.0090	0.0652	0.0000	0.0443	0.2163	0.0000	0.0000	0.0533	0.3039
LR	0.8014	**0.1319**	0.6954	**0.0754**	0.1727	0.5918	0.1558	0.7100	0.7444	0.7466
LR-F	0.8155	0.1247	0.6989	0.0633	0.5469	0.8198	0.1586	0.7269	0.7581	0.7473
LR-B	0.8354	0.1116	0.7057	0.0280	0.2193	0.6357	0.1303	0.7655	0.7868	0.7592
LR-BF	**0.8411**	0.1075	0.7014	0.0269	0.5442	**0.8359**	0.1228	0.7702	**0.7921**	**0.7736**
Ada	0.8042	0.0575	0.6564	0.0102	0.2383	0.4180	0.0729	0.7518	0.7709	0.7088
Ada-F	0.8080	0.0534	0.6774	0.0191	0.4274	0.7852	0.0653	0.7507	0.7738	0.7164

Table 3 details the performance for each machine learning method on each of the PTs in our study. Please see Table 2 for the abbreviation used for each of the PTs. All results are F_1 measures, and we have highlighted in bold the best performing method for each PT.

For each machine learning method, we trained with up to four different feature variations. In all the cases, we considered only Boolean features, either the feature appears in the citation or not:

1. Base method, which includes the text from the Title and Abstract fields. The text has been tokenized, lowercased, and no stemming was applied.
2. Base method plus added text features (-F). For the added text features, we also include the following fields to the default Title and Abstract fields for training: Journal Unique Identifier, Author Affiliations, Author Names, and Grant Agencies. Some of the features rely on either the authors or the institutions to be working on the same type of publications, which might change after some time. The plan is to retrain the learning algorithms to avoid any concept drift.
3. Base method plus bigrams (-B), and
4. Base method plus added text features plus bigrams (-BF).

Due to time constraints, AdaBoostM1 was only trained using the first two variations. We used five different machine learning methods: Modified Huber Loss (Mhl), Hinge Loss (Sgd), Naïve Bayes (NB), Logistic Regression (LR), and AdaBostM1 (Ada). So, in the table under methods, "Mhl-BF" means Modified Huber Loss using bigrams and added text features. We have also highlighted the four PTs (CO, EA, IV, and MA) where we have baseline results from early MTI performance. Even though not directly comparable, the difference in performance is quite significant.

For the four PTs that we have baseline performance information, we can see three have a dramatic improvement with machine learning: *Congresses* improves from 0.3397 to 0.7113 (+**109**%), *English Abstract* improves from 0.0010 to 0.8359 (+**835**%), and *Meta-Analysis* improves from 0.2674 to 0.7742 (+**190**%). Interestingly, *In Vitro* actually has a decrease in performance from 0.1679 to 0.1610 (-**4**%).

Mhl-BF and LR-BF have the best performance from the evaluated methods. These two classifiers have already shown better performance compared to other algorithms in existing work on MeSH indexing [4,5,8,9]. Adding features from the article fields seems to improve the performance compared with using only the Title and Abstract fields. Using bigrams slightly improves the performance.

4 Discussion

Not surprisingly with machine learning, there is no clear winning method that works best for all of the Publication Types, echoing the findings for MeSH indexing [4,5,8,9]. The Logistic Regression (LR) method provides the highest F_1 measures for six of the ten PTs in our study making it the best overall performer. Even within the LR method results, the highest measures come from both the default (LR) and then Base method plus added text features plus bigrams (LR-BF) with a great deal of differences in performance between the two variations. The results for the Modified Huber Loss (Mhl), Hinge Loss (Sgd), and AdaboostM1 (Ada) methods were very close to the results for the LR method and depending on retraining might in some cases perform slightly better than the LR method.

The Naïve Bayes method was far behind all of the other methods. This effect is more dramatic when the ratio of positives is smaller compared to the number of negatives. This has been explained already by Rennie et al. [**Error! Reference source not found.**] and it is due to the imbalance between the classes for which the Naïve Bayes classifier favors the majority class. In addition, this effect is more dramatic with a larger set of dependent features, in which the decision boundary is pushed by the related features favoring the majority class even more.

Case Reports, *Congresses*, *English Abstract*, *Meta-Analysis*, *Randomized Controlled Trial*, and *Review* all have F_1 measures above 0.700 making them promising candidates for future integration into the indexing process. The remaining PTs *Clinical Trial*, *Controlled Clinical Trial*, *Editorial*, and *In Vitro* all have F_1 measures too low for consideration at this time but provide the kernel for further research into improving their performance.

The overall results are promising enough to warrant expanding the experiments to include more PTs to see how they will perform.

If we focus on the 480,631 citations in the 2013 MEDLINE Baseline with a 2012 Publication Date, we can see that several of our high performing PTs were also some of the most frequently used PTs. *Review* (46,808) is fourth, *Case Reports* (27,662) fifth, *English Abstract* (14,208) tenth, and *Randomized Controlled Trial* (11,408) twelfth. By providing accurate Publication Type recommendations to the indexers, we will help make their jobs easier and more efficient.

Two of the PTs intrigued us enough to warrant a deeper study for very different reasons.

English Abstract performed very well (0.8359) in our experiments, but, we could not understand why it did not reach 1.0000. The rule for identifying whether an article is actually an *English Abstract* is very clear, more so than most of the PTs. If an article has a title in brackets (meaning it was translated into English) and contains an ab-

stract, it should receive the *English Abstract* Publication Type. What we found in talking with an indexer is that *English Abstract* is actually not added by the indexer at all. This rule was straightforward enough that there is a program in place to automatically assign this Publication Type to articles before the indexing is released to the MEDLINE/PubMed database. During our false positives error analysis we found that the majority of cases met the definition of *English Abstract*, but, simply did not have the Publication Type assigned and this is very likely the cause of not meeting our goal of 1.0000.

In Vitro on the other hand actually performed worse than our MTI baseline and we wanted to try and find out what might be causing this anomaly. *In Vitro* was designated as a Check Tag when our MTI baseline measure was taken and changed to being a Publication Type shortly thereafter. As a Check Tag, indexers would have used *In Vitro* much differently than as a Publication Type since Check Tags are based on the main topics found in the article and PTs describe the type or genre of the article. This may account for some of the differences in performance, but, there had to be additional reasons for such a low F_1 measure (0.1610) for *In Vitro*. We only used the last three years of MEDLINE in our experiments, so this time period would only include *In Vitro* as a Publication Type. So, we should not be confusing the machine learning algorithms by providing them with contradictory data. What we found in our error analysis was that in almost all of the false negatives that we manually reviewed, the information for designating the article as *In Vitro* was located in the full text of the article, usually in the Methods section, where the authors describe how they performed their research. This fact alone explains the low performance for *In Vitro* and highlights one of the challenges we mentioned earlier (full text versus only using title and abstract) to successfully recommend Publication Types.

5 Conclusion and Future Work

We have evaluated the automatic assignment of PTs to MEDLINE articles based on machine learning, which extends our previous machine learning efforts using MTI. We find that for the majority (6 of 10) of PTs the performance is quite good with F_1 measures above 0.700, while further work is required for the rest of them. The results also show that in addition to the title and abstract text, further information provided from fields in the MEDLINE article result in improved performance. The discussion section shows that feature engineering might provide improved performance, for instance, in the *English Abstract* case.

Future work will involve expanding the experiments to include most of the remaining frequently used PTs to see if we can identify the set of PTs that perform the best and that would provide the most assistance to the indexers. We will also be exploring the

use of openly available full text from PubMed Central[14] to see if the full text would benefit *In Vitro* as well as other poorly performing PTs.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. This work was also partly supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would also like to thank Preeti Kochar a senior indexer at the U.S. National Library of Medicine for her valuable insights into how the Publication Types work from an indexer's perspective.

References

1. Aronson A.R., Bodenreider O., Chang H.F., Humphrey S.M., Mork J.G., Nelson S.J., Rindflesch T.C., Wilbur W.J. (2000). The NLM indexing initiative. Proc AMIA Symp 2000;:17-21.
2. Aronson AR and Lang FM. (2010). An Overview of MetaMap: Historical Perspective and Recent Advances. J Am Med Inform Assoc. 2010 May 1;17(3):229-36.
3. Aronson A.R., Mork J.G., Gay C.W., Humphrey S.M., Rogers W.J. (2004). The NLM Indexing Initiative's Medical Text Indexer. Medinfo 2004;11(Pt 1):268-72
4. Jimeno-Yepes, A., Mork, J.G., Demner-Fushman, D., and Aronson, A.R. (2011c). Automatic algorithm selection for MeSH Heading indexing based on meta-learning. International Symposium on Languages in Biology and Medicine, Singapore, December, 2011.
5. Jimeno-Yepes, Antonio, Mork JG, Demner-Fushman D, Aronson AR. Comparison and combination of several MeSH indexing approaches. AMIA Annual Symposium Proceedings. Vol. 2013. American Medical Informatics Association, 2013.
6. Lin, J., & Wilbur, W. J. (2007). PubMed related articles: a probabilistic topic-based model for content similarity. BMC bioinformatics, 8(1), 423.
7. Bodenreider O, Nelson SJ, Hole WT, and Chang HF. Beyond Synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies. Proc AMIA Symp 1998;:815-9.
8. Jimeno-Yepes, A., Wilkowski, B., Mork, J.G., Demner-Fushman, D., and Aronson, A.R. (2012). MeSH indexing: machine learning and lessons learned. ACM SIGHIT International Health Informatics Symposium, Miami, FL, USA, 2012.
9. Jimeno-Yepes A, Mork J, Demner-Fushman D, Aronson AR. A one-size-fits-all indexing method does not exist: automatic selection based on meta-learning. JCSE, vol. 6, no. 2, pp.151-160, 2012.
10. M.E. Funk and C.A. Reid. Indexing consistency in MEDLINE. Bulletin of the Medical Library Association, 71(2):176, 1983.
11. Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. (C. Nédellec & C. Rouveirol, Eds.)Machine Learning ECML98, 1398(2), 2-7. Springer

[14] http://www.ncbi.nlm.nih.gov/pmc/

12. Zhang, T, Solving large scale linear prediction problems using stochastic gradient descent algorithms. In Proceedings of the Twenty-First International Conference on Machine Learning. ACM, 2004.
13. Yeganova L, Comeau DC, Kim W, Wilbur WJ. Text mining techniques for leveraging positively labeled data. In Proceedings of BioNLP 2011 Workshop (pp. 155-163). Association for Computational Linguistics.
14. Rennie J.D., Shi Rennie J.D., Shih L., Teevan J., Kerger DR. (2003) Tackling the poor assumptions of naive bayes text classifiers. ICML. Vol. 3. 2003.