MTI for Full Text

Semi-automatic indexing for online medical journals

Indexing Initiative,
National Library of Medicine

This report chronicles the investigation of the application of the Medical Text Indexer (MTI) to full text journal articles. It covers all facets of that work: the experiments, the results, and the dead ends.

Introduction

Full Text Indexing is another major area of planned research for the Indexing Initiative. We recognize that our current indexing methods rely only on titles and abstracts, while human indexers base their analysis on the full text of an article. This restriction causes the computer-generated terms to suffer recall errors in comparison to the human assigned document descriptors. Given the increasing availability of machine readable journals, we have begun a full text processing effort.

One approach to full text processing involves submitting all of the text of journal articles to the automatic indexing process. Optimal results are likely to be achieved by addressing those sections of a full-text article which concentrate on the main points of the article. Considerable research in the field of computational linguistics (Lin & Hovy, 1997, for example) is concerned with identifying key topics and sections in a full-text article. Additionally, insights from human indexer practice provide guidance for the automatic methods being developed. For example, in a preliminary study on the effect of key sentences on MetaMap Indexing results, we used the observation of an expert indexer that the last (and sometimes the first) sentence of the introduction of a full journal article often supplies crucial information about how to index the article.

RATIONALE

There are reasons to believe that this research could be a fruitful. The baseline macro precision is 0.52. If the indexing is not limited to the top 25 terms then the recall could

be raised 0.79. This would also raise the f-measure from.0.429 to 0.643, an increase of 51%. Thus if techniques can be developed to accurately select from the MeSH terms found from the text then some portion of this potential could be realized.

The Objective

The overall goal of this research is to improve MTI performance and to determine if that is possible by using full text. Here is a characterization the problem that might be amenable to machine learning and other approaches:

- Given a list of MeSH concepts identified for an article from an online medical journal select and rank the most likely 25 terms to match the main headings from the MEDLINE indexing for that article.

Test Collection

PUBMED CENTRAL

PubMed Central®, a service of the NLM, is a repository of full text articles from online and print published journals. It was selected as the source for the test set since it provides all the articles in a consistent XML format that facilitated processing. PubMed Central provides access to 136 journals. (PubMed Central has 78 journals plus 58 online journals from BioMed Central Ltd. (BMC).)From the 30 journals that are indexed for MEDLINE we selected 17 covering diverse and representative biomedical topics. We chose an issue from September of 2002 for each journal, to assure that the indexing for the journal would be complete. When we found that nearly 15% of the selected articles were coming from one journal, we took a 1 in 10 sample from the issue of the *Proceedings of the National Academy of Science USA* to help maintain the diversity. The resulting collection has 500 articles. To suggest the diversity of the collection the journals include: *Critical Care*, *Genome Research*, and *Plant Physiology*.

TABLE 1. Selected Issues

Journal	Issue	Articles
Antimicrobial Agents and Chemotherapy	Antimicrob Agents Chemother. 2002 Sep;46(9)	65
BMC Biochem	BMC Biochem. 2002;3(1)32	32
BMC Health Services Research	BMC Health Serv Res. 2002 Mar 21;2(1)	22
bmj.com	BMJ 2002 Sep 28;325(7366)	7
Breast Cancer Research	Breast Cancer Res. 2002;4(5)	11
Clinical and Diagnostic Laboratory Immunology	Clin Diagn Lab Immunol. 2002 Sep;9(5)	34
Clinical Microbiology Reviews	Clin Microbiol Rev. 2002 Oct;15(4)	12
Critical Care	Crit Care. 2002 Oct;6(5)	21
Genome Research	Genome Res. 2002 May;12(5)	15
Journal of the American Medical Informatics Association	J Am Med Inform Assoc. 2000 Sep-Oct;7(5)	10
Journal of Bacteriology	J Bacteriol. 2002 Sep;184(17)	36

TABLE 1. Selected Issues

Journal	Issue	Articles
Journal of Clinical Microbiology	J Clin Microbiol. 2002 Sep;40(9)	80
Journal of Virology	J Virol. 2003 Sep;77(17)	60
Learning & Memory	Learn Mem. 2002 Sep-Oct;9(5)	11
Molecular Biology of the Cell	Mol Biol Cell. 2002 Sep;13(9)	30
Nucleic Acids Research	Nucleic Acids Res. 2002 Sep 1;30(17)	33
Plant Physiology	Plant Physiol. 2002 Sep;130(1)	46
Proceedings of the National Academy of Sciences of the United States of America	Proc Natl Acad Sci U S A. 2002 Sep 3;99(18)	8

TABLE 2. Distribution of Journal Categories

Journal Name	Count	Categories												
		Disease	Pharmacology	Bio Chemistry	Genetics	Microbiology	Molecular Biology	Plants	Reviews	Organization	Clinical	Information	Mind	Other
Antimicrobial Agents and Chemotherapy	65		x											
BMC Biochem	32			x										
bmj.com	7	x								x	x			
Breast Cancer Research	11	x									x			
Clinical and Diagnostic Laboratory Immunology	34	x									x			
Clinical Microbiology Reviews	12					x			x		x			
Critical Care	21										x			x
Genome Research	15				x		x							
Journal of the American Medical Informatics Association	10									x		x		
BMC Health Services Research	22													x
Journal of Clinical Microbiology	80					x					x			
Journal of Virology	60													x
Learning & Memory	11												x	
Molecular Biology of the Cell	30						x							
Nucleic Acids Research	33						x							
Plant Physiology	46							x						
Proceedings of the National Academy of Sciences of the United States of America	11									x				x
	500													

Table 1 lists all the journals used in the test collection, the selected issue and the number of articles in that issue. The process of selecting the journals included categorizing all 30 indexed journals and then picking a set that gave broad coverage of domains and views. Table 2 shows the categorization of the final selection.

EVALUATION

To compare different versions of MTI we have chosen to use the F_2 measure, a weighted harmonic mean of recall and precision. Since we want to optimize the performance for the 25 terms we normally recommend, a single measure is preferred. We selected the F_2 measure ($F_\beta = ((\beta^2 + 1)PR)/(\beta^2 P + R)$) over other single value measures because we do not have to treat recall and precision equally. So we use the β=2 version of F measure to reflect that for our users some inappropriate terms can be tolerated if many useful terms are available. The chosen F_2 measure reflects our view that recall is more important as precision. This weighting also ameliorates the built-in handicap of always recommending 25 terms when we know that the normal limit for MeSH terms in MEDLINE is closer to 12. This handicap means that the upper limit for the F2 measure for MTI when recall reaches 100% is 0.822. (For F_1 the limit would be 0.649.)

We compute the F_2 measure for each citation and report the average over all the citations in the experiment since we want to maximize the quality of the individual sets of recommendations rather than the overall performance on the collection. This approach is known as macro-averaging but we average over the documents rather than the classification categories. [YANG 1999]

Establish Baselines

Baselines were established to provide a context for evaluating full text based indexing methods. A test collection of 500 full text articles was created. The title and abstract of those articles were processed normally by MTI to establish the first baseline. The second baseline was the performance of MTI when the body of the article was treated as an abstract and then processed normally, i.e. without any special processing for full text.

BASELINE PERFORMANCE

Here are the results for the two baselines. Note that only 494 of the articles were still in MEDLINE and had indexing.

TABLE 3. MTI Baseline Performance

Text	Identifier	Prec	Recall	Used	F_2 meas	Date
PMC citations	baseline1	.30	.51	7.4	0.438	12/11/03
PMC full text	baseline2 alpha	.26	.55	7.9	0.436	12/16/05
PMC full text	baseline2 beta	.26	.55	7.8	0.432	1/5/04

The baselines show MTI performance on the normal MEDLINE citations, just title and abstract, and its performance on the full-text when the entire article is processed in the

abstract field. Two baselines were run for the full text to verify our approach to testing MTI.

The results from the Related Citations path is unpredictably variable since the TexTool providing matching citations is often rerun on PubMed to reflect newly added citations. Therefore, to insulate our evaluations from this variability, MTI can be run to produce static results from the two indexing paths. This allows enhancements in the post processing of MTI to be tested by applying them to that static data to produce final indexing recommendations. One of the full text baselines was produced using this static data and the other (baseline 2alpha) was produced by MTI running normally. The intent was to show the two forms of MTI processing were identical. As you can see from the results, differences did appear. The two batches took several weeks to run due to errors in MTI processing caused by the extraordinary large size of the articles. These results then illustrate for the need to use the static data approach if small differences in results are to be considered accurate. A sample of 12 articles were processed using both methods on the same day and the results were identical. This gives us the assurance that using the static data does provide representative performance by MTI.

TABLE 4. MTI Baseline 95% Confidence Intervals

	PMC citations	PMC full text
Upper bound	0.44908	0.44140
Micro-averaging Mean	0.43724	0.42973
Lower bound	0.42540	0.41817

The confidence intervals were calculated for these sets of indexing using the individual recommendations, micro-averaging. The full text mean being within the citations confidence intervals shows that there is no significant difference in MTI performance, with or without the full text.

Sections

EXPERIMENTAL COLLECTION

Using the same articles from the PubMed Central test collection, we pulled the sections out and formatted them for MTI processing. These sections were processed by MetaMap Indexing program and by the TexTool for the Related Citations path and the resulting intermediate data was used for the experiments that follow[1].

SECTION CLASSES

The titles from the extracted sections were treated as titles for the pseudo MEDLINE citations processed by MTI. These section titles, which we call headers, were grouped into categories or classes. This clustering was done manually but was based not only on the lexical similarity of two headers, but also based on the patterns of their use that were

1. Within the 500 articles in the test collection, two have been deleted from MEDLINE and four were not indexed because they are book reviews or other publication types that are not indexed.Therefore, the experiments were performed on the remaining 494 articles.

visible is a table of all the sets of headers that structured an articles. These sets of headers are referred to as header frames.

The Headers. The sections extracted from the article structure were these:

- The title and abstract from the fore matter became one section.
- The keywords from the fore matter were a separate section.
- Each of the top level sections and figures or tables from the body of the document were placed in individual sections. The figure and table sections were not included when titles were extracted or header frames were identified.
- From the back matter, that included references, only the Glossary was turned into ac section for processing. These glossaries were usually just dictionaries for abbreviations.

Some section did not have titles but needed to be grouped by their nature or source so some artificial labels were applied to those sections. These labels are pretty obvious: <abstract>, <keywords>, <tblfig>, <none>, <backmatter>.

There are 461 total different headers in the 500 articles. There are 45 above frequency of 1 and 433 appear only once. The top seven are not surprising:

introduction: 414

discussion: 351

results: 347

materials and methods: 323

methods: 50

conclusions: 58

background: 54

The Header Frames. For the 500 articles there were 19 frames with more than 1 occurrence, but more than half of the articles used the most common two frames. Those two frames differed only in the order of the four sections:

introduction|materials and methods|results|discussion: 214

introduction|results|discussion|materials and methods: 50

The Section Classes. Results of manual clustering are in "Appendix A The Section Classes" on page 16. Sections headers were clustered based on their semantic similarity and whether they co-occurred in the test collection.

The 3304 sections were partitioned into thirteen classes formed from 461 distinct headers ranging in frequency from 414 to 29; the 'Other' class has a frequency of 472. The artificial class <backmatter> was merged with actual headers to form a class for sections containing terms and abbreviations. (count below add up to 3313)The remaining artificial headers were put in their own classes and are counted in the thirteen.

<abstract>: 498
<tblfig>: 453
<keywords>:35
<none>: 23

The <none> or no header class contains those sections which were marked as sections but given no title. Often this was because the section was the entire body of a short article such as a letter or editorial. A complete analysis of the role of these sections in their article may be found in Appendix B.

RESULTS

The individual sections where used as the whole representation of the article and the terms recommended by MTI were evaluated. This gave us performance information about each group of sections with the same header and for our section classes. The MTI processing used the normal default settings except that only the MetaMap path was used. (The work with structured abstracts has shown that if was difficult to isolate the contributions of the Related Citations path and the MetaMap when combining the results from the different sections.)

Section Performance: The section performance ranged from many once occurring headers that returned no correct terms to an F_2 measure of 0.61 for the sections labeled:'future perspectives.' Collectively, the sections on average had a precision of 0.18, a recall of 0.30 and an F_2 measure of 0.248. Here are some high scoring headers with more that two occurrences that were not their own classes:

- method: .376
- key messages: .306,
- case report: .303.

More examples can be found in Appendix A.

Section Classes Performance. Table 5 shows the performance results for the sections in each class. The averages are the weighted averages for all the sections in the class. The table is ordered by the relative F_2-measure. Note that captions for the tables and figures of the articles is the only 'section' that is a better source of terms than the abstract.

TABLE 5. Performance by section class - MMI only

Section Class	Section Count	Avg Precision	Avg Recall	Avg F_2 measure
<tblfig>:	64	0.1077	0.7115	0.3175
<abstract+title>:	498	0.2272	0.3452	0.3021
<abstract>	470	0.22	0.34	0.296
introduction:	414	0.1920	0.3412	0.2869
results:	345	0.2016	0.3164	0.2790
discussion:	349	0.1933	0.3138	0.2734
<none>:	23	0.1201	0.3889	0.2574
results and discussion:	28	0.1695	0.2976	0.2542

TABLE 5. Performance by section class - MMI only

Section Class	Section Count	Avg Precision	Avg Recall	Avg F_2 measure
background:	50	0.1742	0.2763	0.2436
<keywords>:	34	0.4585	0.1918	0.2106
materials and methods:	377	0.1364	0.2469	0.2088
conclusions:	80	0.1550	0.2361	0.1961
<other>:	525	0.1037	0.2208	0.1675
abbreviations:	56	0.2329	0.1260	0.1304

There is some variation within the classes. For example ‘materials and methods’ includes ‘method’ at 0.376 and ‘methods’ at 0.187.

The table and figures in that class are not all the tables and figures in the articles. Some articles represented the tables and figures as top level sections; those are evaluated in this class. For other articles the tables and figure are nested within top level sections and those are merely included in the text of the section in which they appeared.

Note that even without the title, the abstract is still a better source of terms than the other sections.

MetaMap Indexing Only

SIMPLE COMBINATION

The first approach treats all of the terms equally from the different sections. The difference from the baseline2 case is that this indexing does not include the contribution of the Related Citations path. So the bench mark for this trial is the abstract with title but no Related Citation terms.

TABLE 6. Performance for simple combination of indexing by section

Sample	Precision	Recall	Avg Used	IM Precision	IM Recall	Avg IM Used	F_2 measure
All Sections	.21	.40	5.77	.09	.64	2.37	.329
Abstract + Title	.23	.35	5.12	.10	.59	2.17	.304
All Sections (2004)	.22	.43	6.14	.09	.64	2.41	.3485
Abstract + Title	.24	.37	.5.46	.11	.59	2.18	.3237

The difference in the F_2 measure for these two samples is only significant at the 90% confidence level. At 95% confidence level the range for the Abstract and title is 0.297 - 0.318 and for the combined sections is 0.317 - 0.338. For the 90% confidence level the ranges are 0.299 - 0.316 and 0.318 - 0.336.

FORWARD BACKWARD SEARCH

We applied step-wise forward and stepwise backward selection to select the best model for combining the terms that found in the different sections of the document. The section classes determined above will be the increments for this process.

MODEL BUILDING

Using a technique often used in machine learning to select text features, we performed a search for the best performing combination of terms from the article sections. The goal was to find the most accurate model of article using the concepts identified by MetaMap and Related Citations. The approach is to take the best performing single section as our seed. Then we process and evaluated the indexing that results from the combination of that section and each of the other section classes. We next take the best performing combination as our base and perform the process again with each of the remaining section classes. This stepwise selection is continued until adding another section class will not improve the performance of the selected section classes. That was the stepwise forward selection. Next we begin stepwise backward selection. We try deselecting in turn each of the selected classes to see if the performance of the collection can be improved. If it does, we repeat as long as successful. Finally, we try the stepwise forward selection process again.

Stepwise Forward Selection Results. After eight rounds of selection, we have the results shown in Table 7. After five or six rounds, performance levels out at 0.351 for

TABLE 7. Stepwise Selection Results - MetaMap only

Selected Section Classes	Number of Matching Terms	F_2 Measure (macro)
<TblFig>	774	0.111
<TblFig> Introduction	2315	0.269
<TblFig> Introduction <Abstract>	2689	0.308
<TblFig>, Introduction, <Abstract>, <Title>	2917	0.337
<TblFig>, Introduction, <Abstract>, <Title>, Results	3014	0.348
<TblFig>, Introduction, <Abstract>, <Title>, Results Discussion	3035	0.350
<TblFig>, Introduction, <Abstract>, <Title>, Results Discussion, <Other>	3069	0.351

TABLE 7. Stepwise Selection Results - MetaMap only

Selected Section Classes	Number of Matching Terms	F_2 Measure (macro)
<TblFig>, Introduction, <Abstract>, <Title>, Results Discussion, <Other>, <None>	3085	0.351
Introduction, <Abstract>, <Title>, Results, Discussion, <Other>, <None>	3082 (3285 - 2004duis)	0.351 (0.3731)

the F_2 measure with only a few additional used terms, terms suggested by MTI that match the MEDLINE indexing.

Stepwise Backward Selection Results. The first round of backward selection did not provide a model that showed improvement in either F_2 measure or total number of matching terms. The best candidate model, shown in the last row of Table 7, resulted from the removal of the <TblFig> class.

The best performing model based on the MetaMap Indexing path alone includes sections from these classes: Introduction, <Abstract>, <Title>, Results, Discussion, <Other>, <None>. It yields a macro F_2 measure of 0.351.

MetaMap and Related Citations

MODEL EXTENSION

The initial model included only indexing recommendations from the MetaMap path of MTI. We now look at adding indexing recommendations from the Related Citations path. There three alternative approaches to using Related Citations. Use citations found by (a) matching on the title and abstract, (b) matching text from each section individually, and (c) matching on the full article text. The baseline system uses approach (a) for Related Citations. Starting with our best model using full text sections, we will apply each of the Related Citations approaches separately, then in combination.

MEDLINE citation. Using the MEDLINE citation as the input for the Related Citations path, we investigated the performance when indexing terms from 1 or more citations selected by Related Citations are considered by MTI. This variant on stepwise selection showed the best results at the maximum available, ten citations. Using the MEDLINE citation (title and abstract) was tried first because the TexTool of the Related Citations path was trained on MEDLINE data, and thus may perform better on that text than on text from the main body of the article. The resulting model raised the number of correct recommendations to 3392 and an F_2 measure of 0.454, a 29% increase. (The possibility of looking at even better results for more than ten citations was considered, but the improvement was beginning to smooth out so the potential improvement did not support he time required to investigate).

After adding Related Citations to the model, stepwise backward selection was applied to the resulting model. Performance was improved through several stages (see results in

Table 8)and then reducing the number of citations from Related Citations was tried. There was no improvement possible.

TABLE 8. Stepwise Selection Results - with Related Citations

Selected Section Classes	Number of Matching Terms	F_2 Measure (macro)
<TblFig>, Introduction, <Abstract>, <Title>, Results Discussion, <Other>	3992	0.454
<TblFig>, <Abstract>, <Title>, Results, Discussion, <Other>	4025	0.459
<TblFig>, <Abstract>, <Title>, Results, <Other>	4043	0.461
<TblFig>, <Abstract>, <Title>, Results,	4044	0.463
<TblFig>, <Abstract>, <Title>, Results, Background	4046	0.464

Continuing with the stepwise process, forward selection was tried again, since four sections had been removed from the model. Addition of the background section produced the minor improvement shown in the last line of Table 8 So for with the use of MetaMap and Related Citations based on MEDLINE citations the best model developed through stepwise selection uses 10 related citations and terms from these sections:

<TblFig>, <Abstract>, <Title>, Results, Background

SEPARATE SECTIONS

Next we investigated the value of adding indexing terms derived by Related Citations based on the text of individual sections. First we investigated the addition of 1-10 citations for each article using the Related Citations data for the sections included in our model. We start back with the best MetaMap only model:

<TblFig>, Introduction, <Abstract>, <Title>, Results, Discussion, <Other>, <None>.

As with the MEDLINE citation based Related Citations, the best performance is achieved when we use all 10 of the available citations for each section

Table 9. shows.the this result in the context of the other major model versions. After three rounds of forward selection, and one round of backwards selection, and a final single round of forward selection this refined model was developed.

<Abstract>, <Title>, <TblFig>, results, conclusions, results and discussion, <None>

The resulting model gives a 0.07 improvement in recall and.0.032 improvement in F_2 measure. This is a 13.7% increase in recall and 7.4% increase in overall performance.

OTHER RESULTS

TABLE 9. Performance for MetaMap and Related Citations

Indexing Model	Precision	Recall	Avg Used	IM Precision	IM Recall	Avg IM Used	F_2 measure
Baseline MTI	.30	.51	7.40	.13	.81	3.06	.438
MetaMap + (Ti,Ab) RC	.28	.56	8.08	.11	.82	3.11	.454
MM + (Ti,Ab) RC refined	.29	.57	8.19	.11	.82	3.12	.464
MM+ RC (all sections)	.29	.58	8.39	.11	.82	3.16	.4700
MM+ RC (common sections)	.29	.57	8.28	.11	.83	3.12	.468
MM + RC refined	.29	.58	8.34	.11	.83	3.12	.4704

Table 10. Performance for MetaMap and Related Citations (2004 duis)

Indexing Model	Precision	Recall	Avg Used	IM Precision	IM Recall	Avg IM Used	F_2 measure	Change
Baseline MTI	.32	.53	7.73	.13	.82	3.08	.457	--
All sections	.27	.57	8.22	.10	.82	3.09	.453	- 0.9%
MetaMap + (Ti,Ab) RC	.29	.59	8.48	.11	.83	3.14	.475	+ 3.9%
MM + (Ti,Ab) RC refined	.30	.60	8.59	.11	.82	3.14	.485	+ 6.1%
MM+ RC (common sections)	.30	.60	8.66	.11	.83	3.13	.488	+ 6.8%
MM + RC refined	.31	.60	8.72	.11	.83	3.13	.491	+7.4%

Performance of Abstracts. For the baseline1 processing we had the abstract and title text. So when we processed abstracts alone with and without titles we are able to make some comparisons. Also since this process was without Related Citations we get a measure of its role in MTI results. The Related Citations contribution is based on the title

TABLE 11. Performance of Abstracts

Abstract +	Precision	Recall	Used	IM Prec.	IM Rec.	IM Used	F_2 measure
Title + RC:	.30	.51	7.40	.13	.81	3.06	.438
NoTitle + RC	.29	.51	7.48	.12	.79	2.99	.429
Title - RC	.23	.25	5.12	.10	.59	2.17	.304
No Title - RC	.22	.34	5.05	.09	.56	2.07	.296

and abstract in all cases.

Dramatic differences here are in recall for results with and without the Related Citations path (+.23 for IM terms). This is reflected in a +0.133 contribution to the F_2 measure from Related Citations. The use of the title has no effect (+.009, +.008) when compared with or without related citations results. (Those differences are less than the confidence interval of +/- .011)

Ranking Function for Related Citations

Using the basic approach of MMI for the frequency factor: for a given section the number of occurrences of a term is divided by the max number of occurrences for any term. The other factor is the average of the current MapScore values from all the occurrences. So MapScore_t = Sum MapScore_ti / n * n/max n

We build a new model for MTI that has a ranking function for the Related Citations path. The function parameters were tuned on regular citations. The new model loses <tblfig> and adds discussion, methods, and introduction.

- S.38.12.m10 4260| 0.4861 --normRC
- .m9 4266| 0.4865 --normRC
- S.30.9.s10 4307| 0.4910
- all text 4132| 0.4718 --normRC
- prod. baseline 3866| 0.4380

The model selected with the ranking function in place does not out perform the previous best model. So the ranking function will not be considered further.

Path Weight

One of the primary parameters of MTI is the weight assigned to indexing path. Normal production settings include 0 for the Trigram method, 2 for Related Citations, and 7 for MetaMap Indexing. Using our best model so far, the ratio was varied between 0.250 and 0.833. The results for F_2 were monotonic on both sides of the maximum which occurs at the default values. Table 12 shows the full set of results from these trials. When the number of citations used for Related Citations was set to 9 there was sufficient degradation of the F_2 measure that additional values were not tried. (14151, 4287 0.4886)(10 is the maximum available with the current data.).

TABLE 12.

Path Weight Settings

Related Citations (pub)	MetaMap Indexing (mmi))	Path Weight Ratio	Number of Recommendations	Terms Matching MEDLINE	F_2 Measure
2	8	.250	14146	4297	0.4900
3	11	.273	14150	4303	0.4906
2	**7**	**.286**	**14151**	**4307**	**0.4910**
2	6	.333	14158	4295	0.4897
3	8	.375	14164	4296	0.4897
3	7	.429	14167	4298	0.4896
3	6	.500	14171	4297	0.4893
4	8	.500	14171	4297	0.4893
4	7	.571	14177	4284	0.4877
5	8	.625	14179	4278	0.4871
4	6	.667	14184	4279	0.4872
5	7	.714	14187	4279	0.4872
5	6	.833	14189	4278	0.4871

Section Weight

The idea behind the next set of experiments is that now that we have a model that maximizes performance when a section is either in or out. Maybe weighting those terms differently or setting the classes currently 0.0 to some low factor might contribute some good terms.

Adjusting the Section Weight value will influence the balance between the Term Weight and the other clustering factors. The first experiment just varies the section parameter between 0.5 and 10. Since the value of the F_2 measure falls off steadily as we move fur-

ther away from the default value of 1.0, more distant values were not investigated. The default remains the best value for this parameter.

TABLE 13. Affect of Section Weight Parameter on Best Model

Section Weight Parameter	Number of Recommendations	Term Matching MEDLINE	F_2 Measure
10.0	14118	4184	0.4779
4.0	14124	4235	0.4838
2.0	14130	4266	0.4866
1.5	14137	4282	0.4883
1.2	14147	4296	0.4900
1.15	14150	4303	0.4906
1.1	14149	4305	0.4908
1.05	14150	4300	0.4902
(default) 1.0	**14151**	**4307**	**0.4910**
0.9	14154	4296	0.4896
0.5	14172	4288	0.4886

SINGLETON HEADERS

The second experiment involves giving second class inclusion to the sections in the <Other> class. This largest of the classes partitioning the sections in the test collection contains the sections from articles that are not organized like a typical research article These did not contribute enough to be added to the model during the stepwise selection process, but may be important to their articles. But as 16% of the sections it seems dangerous to ignore them. So this experiment sets the section weight to a range of values to see if this weighted inclusion would enhance the current model.

Table 14 shows the results of this experiment. Although each of the weights tried for the

TABLE 14. Adding Weight for <Other> Sections

Section Weight Parameter for <Other>	Number of Recommendations	Term Matching MEDLINE	F_2 Measure
(default) 0.0	**14151**	**4307**	**0.4910**
0.1	14330	4312	0.4900
0.25	14329	4312	0.4899
0.5	14322	4315	0.4904
0.9	14318	4308	0.4900

<Other> sections yielded more recommendations matching the MEDLINE indexing the performance was weak and slightly depressed the overall performance.

References

[JOACHIMS, 1998] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning. Springer, 1998.

[JOACHIMS, 2001] T. Joachims, A Statistical Learning Model of Text Classification with Support Vector Machines. Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR), ACM, 2001.

[KO & SEO] .Ko Y, Park J, Seo J. Improving text categorization using the importance of sentences. Information Processing Management, 2004; 40: 65-79.

[LI & ABE 1998] H. Li and N. Abe. Word clustering and disambiguation based on co-occurrence data. In Proceedings of the 17th International Conference on Computational Linguistics, Association for Computational Linguistics. 1998:749-755.

[LIN & HOVY, 1997] Lin, C-Y. and Hovy, E.H. Identifying Topics by Position. Proceedings of the Fifth Conference on Applied Natural Language Processing Conference (ANLP-97), Association for Computational Linguistics, Washington, DC. 1997:283-290.

[YANG 1997 Yang, Y., Pedersen, J.O., A Comparative Study on Feature Selection in Text Categorization, Proc. of the 14th International Conference on Machine Learning ICML97, 1997:412---420,

[YANG 1999] Yiming Yang. An Evaluation of statistical approaches to text categorization. Journal of Information Retrieval. 1999;1(1/2):67-88.

[KO, PARK, SEO] Youngjoong Ko, Jinwoo Park and Jungyun Seo. Improving text categorization using the importance of sentences. Information Processing & Management. 2004 Jan;40(1):65-79

[KEERTHI DECOSTE] S.S. Keerthi and D.M. DeCoste, A modified finite Newton method for fast solution of large scale linear SVMs, Yahoo! Research Labs Tech Report YRL-2004-037.

Download: http://research.yahoo.com/publications/37.pdf

[SCHUEMIE, WEEBER] M. J. Schuemie, M. Weeber , B. J. A. Schijvenaars , E. M. van Mulligen , C. C. van der Eijk , R. Jelier , B. Mons and J. A. Kors. Distribution of information in biomedical abstracts and full-text publicationsBioinformatics 2004 20(16):2597-2604; doi:10.1093/bioinformatics/bth291

Appendix A The Section Classes

This table lists all of the section titles (headers) that are included in each section class. The numbers reflect the rank of the class by individual performance as shown inTable 5. The <other> class has 434 different headers and is represented here by those headers with scores above .50 or count above 1.

TABLE 15.

Section Header	Section Count	Avg Precision	Avg Recall	Avg F2 measure
2. Abstract				
<abstract + title>	498	0.2272	0.3452	0.3021
1. Table or Figure				
<tblfig:	64	0.1077	0.7115	0.3175
4. Introduction				
introduction:	414	0.1920	0.3412	0.2869
3 Discussion				
discussion:	1	0.2857	0.2963	0.2941
discussion	348	0.1930	0.3139	0.2733
5. Results				
results:	344	0.2021	0.3170	0.2796
results and interpretation:	1	0.0385	0.1250	0.0862
6. Methods				
method:	2	0.2148	0.4667	0.3760
experimental procedures:	1	0.2727	0.3750	0.3488
materials and method:	3	0.1984	0.3407	0.2962
materials and methods:	323	0.1367	0.2490	0.2103
methods:	46	0.1279	0.2194	0.1874
scoring methods:	1	0.0417	0.1250	0.0893
other methods tested:	1	0.0385	0.1250	0.0862
7. Conclusions				
summary and conclusions.:	1	0.2222	0.4286	0.3614
conclusion:	17	0.1858	0.2659	0.2210
summary and conclusions:	2	0.2077	0.2822	0.2602
conclusions.:	1	0.1481	0.2222	0.2020
conclusion and outlook:	1	0.0500	0.2500	0.1389
conclusions:	53	0.1518	0.2334	0.1930
summary:	5	0.0715	0.1072	0.0963
8. Abbreviations				
list of abbreviations used:	1	0.2727	0.2308	0.2381

TABLE 15.

Section Header	Section Count	Avg Precision	Avg Recall	Avg F2 measure
list of abbreviations:	3	0.2026	0.1952	0.1964
abbreviations:	35	0.2792	0.1359	0.1402
<backmatter:	17	0.1405	0.0874	0.0923
9. Backgroung				
background:	50	0.1742	0.2763	0.2436
10. Keywords				
<keywords:	34	0.4585	0.1918	0.2106
11. Results and discussion				
results and discussion:	28	0.1695	0.2976	0.2542
12. <Nove>				
<none:	23	0.1201	0.3889	0.2574
13. Other				
future perspectives:	1	0.3103	0.8182	0.6164
implications of the results for breast cancer genetics:	1	0.2903	0.7500	0.5696
concluding remarks:	1	0.2083	1.0000	0.5682
testing of bactec mgit 960 cultures by pcr-reverse cross-blot hybridization assay.:	1	0.3200	0.6667	0.5479
questions:	1	0.4211	0.5714	0.5333
participants, methods, and results:	3	0.1399	0.3660	0.2726
treatment.:	2	0.1400	0.3181	0.2536
comment:	4	0.1297	0.2884	0.2286
imaging studies:	2	0.1813	0.2222	0.1970
laboratory findings:	2	0.0887	0.2777	0.1948
appendix:	3	0.1349	0.1574	0.1364
epidemiology:	3	0.0520	0.2056	0.1242
competing interests:	9	0.0914	0.0819	0.0805
website references:	2	0.1041	0.0667	0.0718
web site reference:	2	0.3750	0.0595	0.0715
web site references:	10	0.0754	0.0490	0.0518
nucleotide sequence accession numbers.:	5	0.1400	0.0385	0.0451
authors' contributions:	24	0.0228	0.0153	0.0163
pre-publication history:	18	0.0222	0.0065	0.0075
supplementary material:	5	0.0000	0.0000	0.0000
14. Title				

Appendix B: Survey of <None> Sections

OBSERVATIONS

A survey of the circumstances of sections in PubMed Central that have not titles found the following cases:

Introductions not labelled. 4 This case the PMC version of the online articles has no title presented for the section, but the online journal version from the publisher has a section title of "Introduction." This was seen in three different journals.

Anonymous first section. 5 One journal had the habit of not labelling the first introductory paragraph. None of these articles had an abstract, but had other labelled sections.

Whole article. 9 Usually letters have one unlabelled section.

Comments/Editiorials. 2 These short articles have other sections but no abstract and start with an unlabelled section.

Errors. 2 Once a section without a header was a subsection with a title (Definitions), but appeared as a top level section in the XML. Another was just part of the Discussion section.

CONCLUSIONS

Primarily the <None> sections are important because they are mostly initial sections in articles without abstracts (20/23) and hence contain critical text for the article.O