MTI for Full Text

Phase 2

This report describes the establishment of baselines for the phase 2 of studies of full text processing by MTI and three initial approaches to improving on the model developed in phase 1.

1.0 Baselines

To provide a stable base for the experiments with full text, the MTI indexing paths were run separately on each of the sections of the full text test collection. The output from each indexing path was saved and subsequently used by MTI for all of the experimental processing. The evaluation in the phase 1 experiments reported in the AMIA paper was based on the human indexing extracted from MEDLINE in December of 2003. Since new work planned for phase 2 would change the way the text from the articles was separated into sections new processing by the indexing paths in MTI was necesary. Since this processing uses the current PubMed database, the processing and evaluation must be done in the current environment. Consequently, moving our experiments to the new environment required establishing new baselines. In addition, the baseline is based on the current production version of MTI, so we prepared an updated version of the experimental, section-handling version of MTI to allow a valid comparison of their results. Finally, to support a current evaluation we extracted from Medline the indexing for the test collection articles to serve at the gold standard. Using that environment new performance baselines were established using the current production version of MTI and the updated, experimental version of MTI.

1.1 Production Baseline

For comparing full text processing to standard, production indexing with MTI, we process the Medline citations for the articles in the test collection using the production version of MTI and the standard options used for DCMS processing. The performance of MTI on that text is shown in Table 1.

1.2 Experimental Baseline

To test whether the development gap between the production and experimental versions of MTI had been closed, the experimental version was run to duplicate the production baseline. The experimental MTI processed the full text articles, but the model used for indexing included just the Title and Abstract. So, processing the same text and running with the same parameters the two systems should produce the same indexing. The experimental baseline is shown in Table 1. The remaining discrepency is still under investigation.

Table 1. Phase 2 Baselines

MTI version	Indexing Model	Precision	Recall	Avg Used	IM Precision	IM Recall	Avg IM Used	F_2 measure	Delta
Production System	Title+Abstract	.32	.53	7.78	.14	.84	3.26	.4576	-
Experimental System	Title+Abstract	.31	.54	7.87	.13	.85	3.28	.4582	+0.13%
Experimental System	Sections30	.30	.59	8.64	.12	.85	3.29	.4846	+5.9%

1.3 Phase 1 Result

The best performing model developed during phase 1 of this exploration of full text use by MTI was identified in the AMIA paper as “Full MTI (refined).” Since the model was built through a stepwise selection process that required many trials, we now refer to this model as Sections30. This name refers to its components and the trial that identified it. For this model MTI only considers terms found in sections belonging to these classes:

<title>,<abstract>, <tblfig>, Results, Results and Discussion, Conclusions,<none>.

This model serves as a baseline for this phase of the work since our goal is to build on the limited success of that model or find techniques that surpass it. However, since there have been many changes in the data used by MTI and in the evaluation environment, it was necessary to rerun the Sections30 model in the current environment and make our future comparisons to its current performance. The final row of Table 1 shows the performance measures for the Sections30 model.

2.0 Number of Related Citations

The experiments in phase 1 used the static data collected from the processing of the individual sections by each of the indexing methods. For the Related Citations path we saved terms from the top 10 citations for each section processed. Experiments varying the number of citations used always found 10 citations to provide the best indexing. Since the data for more than 10 was not available we were unable to determine whether 10 was in fact a maximum. So for phase 2 we collected the 20 top ranked citations for each section on Related Citations path for later use in our experiments. Thus our first experiment was to take the Sections30 model and see what the best performing number of citations would be. The attached figure, “Tuning Number of Citations” shows the range of performance in that dimension. MTI performance now reaches its maximum at 15 related citations with an F_2 measure of 0.4896. This 0.005 increase is welcome but not very significant.(+1.03%)

3.0 Extracting All Table and Figure Text

The titles and captions for tables and figures were treated as separate sections in the phase 1 experiments when they appeared in the XML representation of the full text articles as elements at the same level of the document as the main sections of the article. Since those sections were found to be the most effective source of good index terms, and other researchers have found the text from tables and figures to be important in biomedical articles, we wanted to look at the effect of extracting the titles and captions from all the tables and figures in the articles. We explored two ways of handling the extracted titles and captions from tables and figures. First we collected all the table related text into a single special section and all the figure related data into another. The second approach was to treat the title and caption from each table or figure as a separate section.

These changes dramatically affect the handling of the title and figure text. With the original extraction scheme there were 68 articles with <TblFig> sections in the collection. With the two new schemes that separate out the table and figure text from the sections that contain them, there are 476 articles with <TblFig> sections. The orignal extraction scheme identified only 453 <TblFig> sections. With the 2 sections per article approach there are 1,115 <TblFig> sections processed. With the multiple sections approach there are 2,805 sections. For some perspective, the original extraction scheme had a total 3,304 sections.

3.1 Two Table/Figure Sections

The first row of Table 2 shows the performance of the Sections30 model when applied to the articles with two sections for all tables and figures. The "Delta" column compares each result to the Sections30 baseline. After that basic run with the Section30 model, additional stepwise refinement was conducted and the best number of related citations was determined. The resulting model removes the "Results and Discussion" class of sections from the model and uses 15 related citations. The performance of this Sections44 model is shown in Table 2. Note that this model is 0.0004 below MTI performance with the original extraction and the Sections30 model with 15 related citations.

Table 2. Modified Extraction and Additional Modeling

Extracted Text	Indexing Model	Precision	Recall	Avg Used	IM Precision	IM Recall	Avg IM Used	F_2 measure	Delta
2 Tb/lFig sections	Sections30	.30	.60	8.68	.12	.85	3.30	.4857	+0.23%
2 Tbl/Fig sections	Sections44	.30	.60	8.74	.12	.85	3.32	.4892	+0.95%
Multiple Tbl/Fig Sections	Sections30	.30	.59	8.64	.12	.84	3.28	.4834	-0.26%
Multiple Tbl/Fig Sections	Sections30 + 18cits	.30	.60	8.70	.12	.85	3.31	.4870	+0.50%

3.2 Multiple Table/Figure Sections

The text extracted in this approach places the title and caption of each table or figure appearing in the full text article in a separate section. The effect of this difference in the MTI processing is that each section gets its own set of related citations, dramatically increasing the amount of input from the Related Citations indexing method. The number of <TblFig> sections increases six fold with this extraction approach. Again after a basic run with the Section30 model, stepwise refinement was applied and the number of related citations tuned. The initial results were less than the results with just two sections, so subsequent refinement was abandoned. The basic results and the Sections30 model with 18 related citations are shown in Table 2.

4.0 Section Titles

The original extraction of the section text used the title of each section just as a label to identify the section. Later during modeling, the label determined to what section class the section belonged and how its text would be used or not use by MTI. Handled like this, the section titles were not processed by MMI or Related Citations. For sections with titles like "Results" or "Introduction" this was clearly not significant. However, for sections in the <Other> class with titles such as "Implications of the Results for Breast Cancer Genetics" that are often content bearing, the text in those section titles might be a useful source of indexing terms. (The <Other> class contains all those sections whose headers were eiher unique or had a very low frequency.To check out this premise, an version of the extracted sections put the title of each section in the title(TI) field of the pseudo-Medline citations processed by MTI. This was done in combination with each of the table and figure approaches.

The basic run processing the titles with the two sections for the table and figure text yielded an F_2 measure of 0.4857 and the related citations tuning selected 15 citations for a F_2 measure of 0.4886. The basic run showed no change and the 15 related citations run lowered the F_2 measure by -0.0002. This would be expected because the <Other> class does not belong to the Sections30 model. However, adding the <Other> class did not improve the model. Apparently the standard section titles create more noise than the term rich titles add helpful terms.

When combined with the multiple table/figure sections approach, the basic run with the Sections30 model yields F_2 of 0.4835 (+0.0001) and the optimal number of 14 related citations yields an F_2 of 0.4870. As expected this score matches the multiple section approach, but the inclusion of the <Other> class in the model, actually reduced the F_2 measure (0.4856). Table 3 compares the MTI performance of the various models with and without using the title data. These results suggest that using the title text in this way does not help MTI performance on full text articles.

TABLE 3. Processing Section Title Text

Extracted Text	Model	F_2 Measure	Delta
2 Tbl/Fig Sections	Sections30	0.4857	--
2 Tbl/Fig Sections + Titles	Sections30	0.4857	+0.00%
2 Tbl/Fig Sections	Sections30 + 15 cits	0.4888	--
2 Tbl/Fig Sections + Titles	Sections30 + 15 cits	0.4886	-0.04%
2 Tbl/Fig Sections + Titles	Sect30 + <Other> +15	0.4865	-0.47%
Multiple Sections	Sections30	0.4834	--
Multiple Sections + Titles	Sections30	0.4835	+0.83%
Multiple Sections	Sections30 + 14 cits	0.4870	--
Multiple Sections + Titles	Sections30 + 14 cits	0.4870	+0.00%
Multiple Sections + Titles	Sect30 + <Other> +14	0.4856	-0.29%

5.0 Summary

The best performing models and their F_2 scores for each of the extraction strategies studied are shown in Table 4. Although each new extraction was able to improved over the original Sections30 baseline result of 0.4846, the improvement due to the change in the extracted text and the tuning of the number of related citations was always less than the +0.0050 achieved through tuning the Sections30 model. Since the difference in the best model with the original extraction and the best model with the two table and figure sections extraction is so small (.0004) it might be helpful to evaluate these two models using a four-fold cross-validation technique.

TABLE 4. Best Performing Models

Extracted Text	Model	F_2 Measure	Delta
Original	Sections30	0.4846	--
Original	Sections30 + 15 cits	0.4896	+1.03%
2 Tbl/Fig Sections	Sections44	0.4892	+0.95%
Multiple Tbl/Fig Sections	Sections30 + 18 cits	0.4870	+0.50%
2 sections + Titles	Sections30 + 15 cits	0.4886	+0.83%
Multiple Sections + Titles	Sections30 + 14 cits	0.4870	+0.50%

Thus our current optimal model is the Sections30 model with the use of 15 related citations. When compared to the production baseline performance of 0.4576, the current optimal model achieves 0.4896, an overall improvement of 7.0%.