From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches

Alan R. Aronson[1], **Olivier Bodenreider**[1], **Dina Demner-Fushman**[1], **Kin Wah Fung**[1], **Vivian K. Lee**[1,2], **James G. Mork**[1], **Aurélie Névéol**[1], **Lee Peters**[1], **Willie J. Rogers**[1]

[1]Lister Hill Center
National Library of Medicine
Bethesda, MD 20894
{alan, olivier, demnerd, kwfung, mork, neveola, peters, wrogers}
@nlm.nih.gov

[2]Vanderbilt University
Nashville, TN 37235
vivian.lee@vanderbilt.edu

Abstract

This paper describes the application of an ensemble of indexing and classification systems, which have been shown to be successful in information retrieval and classification of medical literature, to a new task of assigning ICD-9-CM codes to the clinical history and impression sections of radiology reports. The basic methods used are: a modification of the NLM Medical Text Indexer system, SVM, k-NN and a simple pattern-matching method. The basic methods are combined using a variant of stacking. Evaluated in the context of a Medical NLP Challenge, fusion produced an F-score of 0.85 on the Challenge test set, which is considerably above the mean Challenge F-score of 0.77 for 44 participating groups.

1 Introduction

Researchers at the National Library of Medicine (NLM) have developed the Medical Text Indexer (MTI) for the automatic indexing of the biomedical literature (Aronson et al., 2004). The unsupervised methods within MTI were later successfully combined with machine learning techniques and applied to the classification tasks in the Genomics Track evaluations at the Text Retrieval Conference (TREC) (Aronson et al., 2005 and Demner-Fushman et al., 2006). This fusion approach consists of using several basic classification methods with complementary strengths, combining the results using a modified ensemble method based on stacking (Ting and Witten, 1997).

While these methods have shown reasonable performance on indexing and retrieval tasks of biomedical articles, it remains to be determined how they would perform on a different biomedical corpus (e.g., clinical text) and on a different task (e.g., coding to a different controlled vocabulary). However, except for competitive evaluations such as TREC or BioCreAtIvE, corpora and gold standards for such tasks are generally not available, which is a limiting factor for such studies. For a survey of currently available corpora and developments in biomedical language processing, see Hunter and Cohen, 2006.

The Medical NLP Challenge[1] sponsored by a number of groups including the Computational Medicine Center (CMC) at the Cincinnati Children's Hospital Medical Center gave us the opportunity to apply our fusion approach to a clinical corpus. The Challenge was to assign ICD-9-CM codes (International Classification of Diseases, 9th Revision, Clinical Modification)[2] to clinical text consisting of anonymized clinical history and impression sections of radiology reports.

The Medical NLP Challenge organizers distributed a training corpus of almost 1,000 of the anonymized, abbreviated radiology reports along with

[1] See www.computationalmedicine.org/challenge/.
[2] See www.cdc.gov/nchs/icd9.htm.

BioNLP 2007: Biological, translational, and clinical language processing, pages 105–112,
Prague, June 2007.

gold standard ICD-9-CM assignments for each report obtained via a consensus of three independent sets of assignments. The primary measure for the Challenge was defined as the balanced F-score, with a secondary measure being cost-sensitive accuracy. These measures were computed for submissions to the Challenge based on a test corpus similar in size to the training corpus but distributed without gold standard code assignments.

The main objective of this study is to determine what adaptation of the original methods is required to code clinical text with ICD-9-CM, in contrast to indexing and retrieving MEDLINE®. Note that an earlier study (Gay et al., 2005) showed that only minor adaptations were required in extending the original model to full-text biomedical articles. A secondary objective is to evaluate the performance of our methods in this new setting.

2 Methods

In early experimentation with the training corpus provided by the Challenge organizers, we discovered that several of the training cases involved negated assertions in the text and that deleting these improved the performance of all basic methods being tested. For example, “no pneumonia” occurs many times in the impression section of a report, sometimes with additional context. Section 2.1 describes the process we used to remove these negated expressions; section 2.2 consists of descriptions of the four basic methods used in this study; and section 2.3 defines the fusion of the basic methods to form a final result.

2.1 Document Preparation

The NegEx program (Chapman et al., 2001a and 2001b, and Goldin and Chapman, 2003), which discovers negated expressions in text, was used to find negated expressions in the training and test corpora using a dictionary generated from concepts from the 2006AD version of the UMLS® Metathesaurus® (excluding the AMA vocabularies). A table containing the concept unique identifier (CUI) and English string (STR with LAT=‘ENG’) was extracted from the main concept table, MRCON, and was used as input to NegEx to generate a dictionary that was later used as the universe of expressions which NegEx could find to be negated in the target corpora. (See the Appendix for examples of the input and output to this process.)

The XML text of the training and test corpora was converted to a tree representation and then traversed, operating on one radiology report at a time. The clinical history and impression sections of each report were tokenized to allow whitespace to be separated from the punctuation, numbers and alphabetic text. The concepts from the UMLS were tokenized in the same way, to allow the concepts found by NegEx to be aligned with the text. The negation phrases discovered by NegEx were also tokenized to find the appropriate negation phrase preceding or trailing the target concept. Using the location information obtained by matching the set of one or more target concepts and the associated negation phrase, the overlapping concept spans were merged and the span for the negation phrase and the outermost negated concept was removed. Any intervening concepts associated with the same negation phrase were removed, too. The abbreviated tree representation was then re-serialized back into XML.

As an example of our use of NegEx, consider the report with clinical history “13-year 2-month - old female evaluate for cough.” and impression “No focal pneumonia.” After removal of negated text, the clinical history becomes “13-year 2-month - old female”, and the discussion is empty.

2.2 Basic Methods

The four basic methods used for the Medical NLP Challenge are MTI (a modification of NLM’s Medical Text Indexer system), SVM (Support Vector Machines), k-NN (k Nearest Neighbors) and Pattern Matching (a simple, pattern-based classifier). Each of these methods is described here. Note that the MTI method uses a “Restrict to ICD-9-CM” algorithm that is described in the next section.

MTI. The original Medical Text Indexer (MTI) system, shown in Figure 1, consists of an infrastructure for applying alternative methods of discovering MeSH® headings for citation titles and abstracts and then combining them into an ordered list of recommended indexing terms. The top portion of the diagram consists of two paths, or methods, for creating a list of recommended indexing terms: MetaMap Indexing and PubMed® Related Citations. The MetaMap Indexing path actually computes UMLS Metathesaurus concepts, which are passed to the Restrict to MeSH process (Bodenreider et al., 1998). The results from each path are weighted and combined using Post-Processing, which also refines the results to conform to NLM indexing policy. The system is highly parameterized not only by path weights but also by several parameters specific to the Restrict to MeSH and Post-Processing processes.

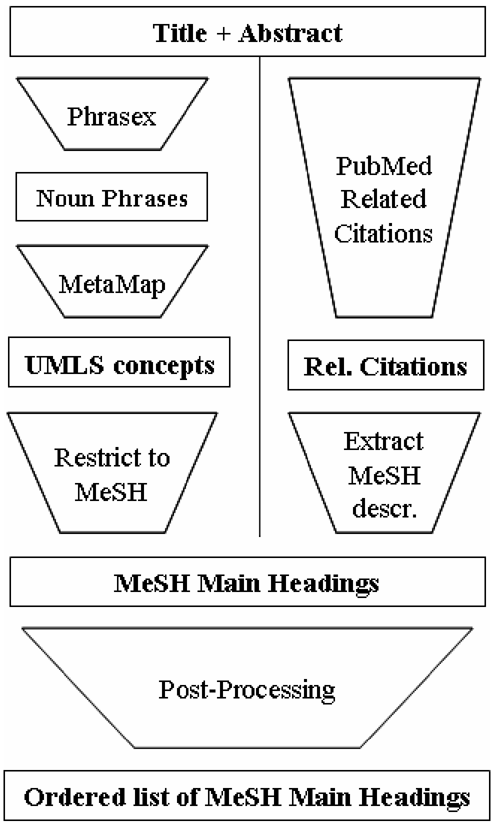

Figure 1: Medical Text Indexer (MTI) System

For use in the Challenge, the Medical Text Indexer (MTI) program itself required few adaptations. Most of the changes involved the environment from which MTI obtains the data it uses without changing the normal parameter settings. We also added a further post-processing component to filter our results.

For the environment, we replaced MTI's normal "Restrict to MeSH" algorithm with a "Restrict to ICD-9-CM" algorithm, described below, in order to map UMLS concepts to ICD-9-CM codes instead of MeSH headings. We also trained the PubMed Related Citations component, TexTool (Tanabe and Wilbur, 2002), on the Medical NLP Challenge training data instead of the entire MEDLINE/PubMed database as is the case for normal MTI use at NLM. For both of these methods, we used the actual ICD-9-CM codes to mimic UMLS CUIs used internally by MTI.

To create the new training data for the TexTool (Related Citations), we reformatted the Medical NLP Challenge training data into a pseudo-MEDLINE format using the "doc id" component as the PMID, the "CLINICAL_HISTORY" text component for the Title, the "IMPRESSION" text component for the Abstract, and all of the "CMC_MAJORITY" codes as MeSH Headings (see Figure 2). This provided us with direct ICD-9-CM codes to work with instead of MeSH Headings.

```
<doc id="97663756" type="RADIOLOGY_REPORT">
  <codes>
    <code origin="CMC_MAJORITY" type="ICD-9-
CM">780.6</code>
    <code origin="CMC_MAJORITY" type="ICD-9-
CM">786.2</code>
    <code origin="COMPANY3" type="ICD-9-
CM">786.2</code>
    <code origin="COMPANY1" type="ICD-9-
CM">780.6</code>
    <code origin="COMPANY1" type="ICD-9-
CM">786.2</code>
    <code origin="COMPANY2" type="ICD-9-
CM">780.6</code>
    <code origin="COMPANY2" type="ICD-9-
CM">786.2</code>
  </codes>
  <texts>
    <text origin="CCHMC_RADIOLOGY"
type="CLINICAL_HISTORY">Cough and fever.</text>
    <text origin="CCHMC_RADIOLOGY"
type="IMPRESSION">Normal radiographic appear-
ance of the chest, no pneumonia.</text>
  </texts>
</doc>
```

```
PMID- 97663756
TI  - Cough and fever.
AB  - Normal radiographic appearance of the
chest, no pneumonia.
MH  - Fever (780.6)
MH  - Cough (786.2)
```

Figure 2: XML Medical NLP Training Data modified to pseudo-ASCII MEDLINE format

Within MTI we also utilized an experimental option for MetaMap (Composite Phrases), which provides a longer UMLS concept match than usual. We did not use the following: (1) UMLS concept-specific checking and exclusion sections; and (2) the MeSH Subheading generation, checking, and removal elements, since they were not needed for this Challenge. We then had MTI use the new Restrict to ICD-9-CM file and the new TexTool to generate its results.

Restrict to ICD-9-CM. The mapping of every UMLS concept to ICD-9-CM developed for the Medical NLP Challenge is an adaptation of the original mapping to MeSH, later generalized to any target vocabulary (Fung and Bodenreider, 2005). Based on the UMLS Metathesaurus, the mapping utilizes four increasingly aggressive techniques: synonymy, built-in mappings, hierarchical mappings and associative mappings. In order to comply with coding rules in ICD-9-CM, mappings to non-leaf codes are later resolved into leaf codes.

Mappings to ICD-9-CM are identified through ***synonymy*** when names from ICD-9-CM are included in the UMLS concept identified by MetaMap. For example, the ICD-9-CM code 592.0 *Calculus of kidney* is associated with the UMLS concept C0392525 *Nephrolithiasis* through synonymy.

Built-in mappings are mapping relations between UMLS concepts implied from mappings provided by source vocabularies in the UMLS. For example, the UMLS concept C0239937 *Microscopic hematuria* is mapped to the concept C0018965 (which contains the ICD-9-CM code 599.7 *Hematuria*) through a mapping provided by SNOMED CT.

In the absence of a mapping through synonymy or built-in mapping, a ***hierarchical mapping*** is attempted. Starting from the concept identified by MetaMap, a graph of ancestors is built by first using its parent concepts and broader concepts, then adding the parent concepts and broader concepts of each concept, recursively. Semantic constraints (based on semantic types) are applied in order to prevent semantic drift. Ancestor concepts closest to the MetaMap source concept are selected from the graph. Only concepts that can be resolved into ICD-9-CM codes (through synonymy or built-in mapping) are selected. For example, starting from C0239574 *Low grade pyrexia*, a mapping is found to ICD-9-CM code 780.6 *Fever*, which is contained in the concept C0015967, one of the ancestors of C0239574.

The last attempt to find a mapping involves not only hierarchical, but also associative relations. Instead of starting from the concept identified by MetaMap, ***associative mappings*** explore the concepts in associative relation to this concept. For example, the concept C1458136 *Renal stone substance* is mapped to ICD-9-CM code 592.0 *Calculus of kidney*.

Finally, when the identified ICD-9-CM code was not a leaf code (e.g., 786.5 *Chest pain*), we remapped it to one of the corresponding leaf codes in the training set where possible (e.g., 786.50 *Unspecified chest pain*).

Of the 2,331 UMLS concepts identified by MetaMap in the test set after freezing the method, 620 (27%) were mapped to ICD-9-CM. More specifically, 101 concepts were mapped to one of the 45 target ICD-9-CM codes present in the training set. Of the 101 concepts, 40 were mapped through synonymy, 11 through built-in mappings, 40 through hierarchical mapping and 10 through associative mapping.

After the main MTI processing was completed, we applied a post-processing filter, restricting our results to the list of 94 valid combinations of ICD-9-CM codes provided in the training set (henceforth referred to as allowed combinations) and slightly emphasizing MetaMap results. Examples of the post-processing rules are:

- If MTI recommended 079.99 (Unspecified viral infection in conditions…) via either MetaMap or Related Citations, use 079.99, 493.90 (Asthma, unspecified type…), and 780.6 (Fever) for indexing. This is the only valid combination for this code based on the training corpus.
- Similarly, if MTI recommended “Enlargement of lymph nodes” (785.6) via the MetaMap path with a score greater then zero, use 785.6 and 786.2 (Cough) for indexing.

The best F-score (F = 0.83) for the MTI method was obtained on the training set using the negation-removed text. This was a slight improvement over using the original text (F = 0.82).

SVM. We utilized Yet Another Learning Environment[3] (YALE), an open source application developed for machine learning and data mining, to determine the data classification performance of support vector machine (SVM) learning on the

[3] See http://rapid-i.com.

training data. To prepare the Challenge data for analysis, we removed all stop words and created feature vectors for the free text extracted from the "CLINICAL_HISTORY" and "IMPRESSION" fields of the records. Since both the training and test Challenge data had a known finite number of individual ICD-9-CM labels (45) and distinct combinations of ICD-9-CM labels (94), the data was prepared both as feature vectors for 45 individual labels as well as a model with 94 combination labels. In addition, the feature vectors were created using both simple term frequency as well as inverse document frequency (IDF) weighting, where the weight is (1+log(term frequency))*(total documents/document frequency). There were thus a total of four feature vector datasets: 1) 45 individual ICD-9-CM labels and simple term frequency, 2) 45 ICD-9-CM labels and IDF weighting, 3) 94 ICD-9-CM combinations and simple term frequency, and 4) 94 ICD-9-CM combinations and IDF weighting.

The YALE tool encompasses a number of SVM learners and kernel types. For the classification problem at hand, we chose the C-SVM learner and the radial basis function (rbf) kernel. The C-SVM learner attempts to minimize the error function

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \quad ,$$

$$\gamma_i(w^T\varphi(x_i)+b) \geq 1-\xi_i \ and \ \xi_i \geq 0, i=1,\ldots,N$$

where w is the vector of coefficients, b is a constant, φ is the kernel function, x are the independent variables, and ξ_i are parameters for handling the inputs. C > 0 is the penalty parameter of the error function. The rbf kernel is defined as $K(x, x') = \exp(-\gamma\,|x - x'|^2)$, $\gamma > 0$ where γ is a kernel parameter that determines the rbf width. We ran cross-validation experiments using YALE on all training datasets and varying C (10, 100, 1000, 10000) and γ (0.01, 0.001, 0.0001, 0.00001) to determine the optimal C and γ combination. The cross-validation experiments generated classification models that were then applied to the complete training datasets to analyze the performance of the learner. The 94 ICD-9-CM combination and simple term frequency dataset with C = 10000 and γ = 0.01 had the best F-score at 0.86. The best F-score for the 94 ICD-9-CM combination and IDF weight dataset was 0.79, where C = 0.001 and γ = 10000.

Further preprocessing the training dataset by removing negated expressions was found to improve the best F-score from 0.86 to 0.87. The C = 10000 and γ = 0.01 combination was then applied to the test dataset, which was preprocessed to remove negation and stop words and transformed to a feature vector using 94 ICD-9-CM combinations and simple term weighting. The predicted ICD-9-CM classifications and confidence of the predictions for each clinical free text report were output and later combined with other methods to optimize the accuracy and precision of our ICD-9-CM classifications.

k-NN. The Challenge training set was used to build a k-NN classifier. The k-NN classification method works by identifying, within a labelled set, documents similar to the document being classified, and inferring a classification for it from the labels of the retrieved neighbors.

The free text in the training data set was processed to obtain a vector-space representation of the patient reports.

Several methods of obtaining this representation were tested: after stop words were removed, simple term frequency and inverse document frequency (IDF) weighting were applied alternatively. A higher weight was also given to words appearing in the history portion of the text (*vs.* impression). Eventually, the most efficient representation was obtained by using controlled vocabulary terms extracted from the free text with MetaMap.[4] Further processing on this representation of the training data showed that removing negated portions of the free text improved the results, raising the F-score from 0.76 to 0.79.

Other parameters were also assessed on the training data, such as the number of neighbors to use (2 was found to be the best *vs.* 5, 10 or 15) and the restriction of the ICD-9-CM predictions to the set of 94 allowed combinations. When the prediction for a given document was not within the set of allowed 94 combinations, an allowed subset of the ICD-9-CM codes predicted was selected based on the individual scores obtained for each ICD-9-CM code.

The best F-score (F = 0.79) obtained on the training set used the MetaMap-based representa-

[4] Note that this use of MetaMap is independent of its inclusion as a component of MTI.

tion with simple frequency counts on the text with negated expressions removed. ICD-9-CM predictions were obtained from the nearest neighbors and restricted to one of the 94 allowed combinations.

Pattern Matching. We developed a pattern-matching classifier as a baseline for our more sophisticated classification methods. A list of all UMLS string representations for each of 45 codes (including synonyms from source vocabularies other than ICD-9-CM) was created as described in the MTI section above. The strings were then converted to lower case, punctuation was removed, and strings containing terms unlikely to be found in a clinical report were pruned. For example, *Abdomen NOS pain* and *Abdominal pain (finding)* were reduced to *abdominal pain*. For the same reasons, some of the strings were relaxed into patterns. For example, it is unlikely to see *PAIN CHEST* in a chart, but very likely to find *pain in chest*. The string, therefore, was relaxed to the following pattern: *pain.*chest*. The text of the clinical history and the impression fields of the radiology reports with negated expressions removed (see Section 2.2) was broken up into sentences. Each sentence was then searched for all available patterns. A corresponding code was assigned to the document for each matched pattern. This pattern matching achieved F-score = 0.79 on the training set. To reduce the number of codes assigned to a document, a check for allowed combinations was added as a post-processing step. The combination of assigned codes was looked up in the table of allowed codes. If not present, the codes were reduced to the combination of assigned codes most frequently occurring in the training set. This brought the F-score up to 0.84 on the training data. As the performance of this classifier was comparable to other methods, we decided to include these results when combining the predictions of the other classifiers.

2.3 Fusion of Basic Methods: Stacking

Experience with ad hoc retrieval tasks in the TREC Genomics Track has shown that combining predictions of several classifiers either significantly improves classification results, or at least provides more consistent and stable results when the training data set is small (Aronson et al., 2005). We therefore experimented with stacking (Ting and Witten, 1997), using a simple majority vote and a union of all assigned codes as baselines. The predictions of base classifiers described in the previous section were combined using our re-implementation of the stacked generalization proposed by Ting and Witten.

3 Results

Table 1 shows the results obtained for the training set. The best stacking results were obtained using predictions of all four base classifiers on the text with deleted negated expressions and with checking for allowed combinations. We retained all final predictions with probability of being a valid code greater than 0.3. Checking for the allowed combinations for the ensemble classifiers degraded the F-score significantly.

Classifier	F-score
MTI	0.83
SVM	0.87 (x-validation)
k-NN	0.79 (x-validation)
Pattern Matching	0.84
Majority	0.82
Stacking	**0.89**

Table 1: Training results for each classifier, the majority and stacking

Since stacking produced the best F-score on the training corpus and is known to be more robust than the individual classifiers, the corresponding results for the test corpus were submitted to the Challenge submission website. The stacking results for the test corpus achieved an F-score of 0.85 and a secondary, cost-sensitive accuracy score of 0.83. For comparison purposes, 44 Challenge submissions had a mean F-score of 0.77 with a maximum of 0.89. Our F-score of 0.85 falls between the 70th and 75th percentiles.

4 Discussion

It is significant that it was fairly straightforward to port various methods developed for ad hoc MEDLINE citation retrieval, indexing and classification to the assignment of codes to clinical text. The modifications to MTI consisted of replacing Restrict to MeSH with Restrict to ICD-9-CM, training the Related Citations method on clinical text and replacing MTI's normal post-processing with a much simpler version. Preprocessing the text using

NegEx to remove negated expressions was a further modification of the overall approach.

It is noteworthy that a simple pattern-matching method performed as well as much more sophisticated methods in the effort to fuse results from several methods into a final outcome. This unexpected success might be explained by the following limitations of the Challenge.

Possible limitations on the extensibility of the current research arise from two observations: (1) the Challenge cases were limited to two relatively narrow topics, cough/fever/pneumonia and urinary/kidney problems; and (2) the clinical text was almost error-free, a situation that would not be expected in the majority of clinical text. It is possible that these conditions contributed to the success of the pattern-matching method but also caused anomalous behavior, such as the fact that simple frequency counts provided a better representation than IDF for the SVM and k-NN methods.

Finally, as a result of low confidence in the ICD-9-CM code assignment, no codes were assigned to 29 records in the test set. It is worthwhile to explore the causes for such null assignments. One of the reasons for low confidence could be the aggressive pruning of the text by the negation algorithm. For example, after removal of negated text in the sample report given in section 2.1, the only remaining text is "13-year 2-month - old female" from the clinical history field; this provided no evidence for code assignment. Secondly, in some cases the original text was not sufficient for confident code assignment. For example, for the document with clinical history "Bilateral grade 3." and impression "Interval growth of normal appearing Kidneys", no code was assigned by the SVM, k-NN, or pattern-matching classifiers. Code 593.70 corresponding to the UMLS concept *Vesicoureteral reflux with reflux nephropathy, unspecified or without reflux nephropathy* was assigned by MTI with a very low confidence, which was not sufficient for the final assignment of the code. The third reason for assigning no code to a document was the wide range of assignments provided by the base classifiers. For example, for the following document: "CLINICAL_HISTORY: 3-year - old male with history of left ureteropelvic and ureterovesical obstruction. Status post left pyeloplasty and left ureteral reimplantation. IMPRESSION: 1. Stable appearance and degree of hydronephrosis involving the left kidney. Stable urothelial thickening. 2. Interval growth of kidneys, left greater than right. 3. Normal appearance of the right kidney with interval resolution of right urothelial thickening." MTI assigned codes 593.89 *Other specified disorders of kidney and ureter* and 591 *Hydronephrosis*. Codes 593.70 *Vesicoureteral reflux with reflux nephropathy, unspecified or without reflux nephropathy* and 753.3 *Double kidney with double pelvis* were assigned by the k-NN classifier. Pattern matching resulted in assignment of code 591 with fairly low confidence. No code was assigned to this document by the SVM classifier. Despite failing to assign codes to these 29 records, the conservative approach (using threshold) resulted in better performance, achieving F-score 0.85 compared to F-score 0.80 when all 1,634 codes assigned by the base classifiers were used.

5 Conclusion

We are left with two conclusions. First, this research confirms that combining several complementary methods for accomplishing tasks, ranging from ad hoc retrieval to categorization, produces results that are better and more stable than the results for the contributing methods. Furthermore, we have shown that the basic methods employing domain knowledge and advanced statistical algorithms are applicable to clinical text without significant modification. Second, although there are some limitations of the current Challenge test collection of clinical text, we appreciate the efforts of the Challenge organizers in the creation of a test collection of clinical text. This collection provides a unique opportunity to apply existing methods to a new and important domain.

Acknowledgements

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine and by appointments of Aurélie Névéol and Vivian Lee to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

The authors gratefully acknowledge the many essential contributions to MTI, especially W. John Wilbur for the PubMed Related Citations indexing method, and Natalie Xie for adapting TexTool (an interface to Related Citations) for this paper.

References

Aronson AR, Demner-Fushman D, Humphrey SM, Lin J, Liu H, Ruch P, Ruiz ME, Smith LH, Tanabe LK, Wilbur WJ. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. Proc TREC 2005, 36-45.

Aronson AR, Mork JG, Gay CW, Humphrey SM and Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004: 268-72.

Bodenreider O, Nelson SJ, Hole WT and Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp 1998: 815-9.

Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan B. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp. 2001a:105-9.

Chapman WW, Bridewell W, Hanbury P, Cooper GF and Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001b;34:301-10.

Demner-Fushman D, Humphrey SM, Ide NC, Loane RF, Ruch P, Ruiz ME, Smith LH, Tanabe LK, Wilbur WJ and Aronson AR. Finding relevant passages in scientific articles: fusion of automatic approaches vs. an interactive team effort. Proc TREC 2006, 569-76.

Fung KW and Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc 2005: 266-70.

Gay CW, Kayaalp M and Aronson AR. Semi-automatic indexing of full text biomedical articles. AMIA Annu Symp Proc. 2005:271-5.

Goldin I and Chapman WW. Learning to detect negation with 'not' in medical texts. Proc Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR, 2003.

Hunter L and Cohen KB. Biomedical language processing: what's beyond PubMed? Mol Cell. 2006 Mar 3;21(5):589-94.

Tanabe L and Wilbur WJ. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, Aug 2002; 18: 1124 –32.

Ting WK and Witten I. 1997. Stacking bagged and dagged models. 367-375. Proc. of ICML'97. Morgan Kaufmann, San Francisco, CA.

Appendix

A sample of the input to NegEx for dictionary generation:

```
C0002390 pneumonitis, allergic interstitial
C0002390 allergic interstitial pneumonitis, nos
C0002390 extrinsic allergic bronchiolo alveolitis
C0002390 extrinsic allergic bronchiolo alveolitis, nos
C0002390 hypersensitivity pneumonia
C0002390 hypersensitivity pneumonia, nos
C0002390 eaa  extrinsic allergic alveolitis
C0002390 allergic extrinsic alveolitis nos (disorder)
C0002390 extrinsic allergic alveolitis (disorder)
C0002390 hypersensitivity pneumonitis nos (disorder)
```

A sample of the dictionary generated by NegEx for later use in detecting negated expressions:

```
C0002098 hypersensitivity granuloma (morphologic abnormality
C0151726 hypersensitivity injection site
C0020517 hypersensitivity nos
C0429891 hypersensitivity observations
C0002390 hypersensitivity pneumonia
C0002390 hypersensitivity pneumonia, nos
C0002390 hypersensitivity pneumonitides
C0005592 hypersensitivity pneumonitides, avian
C0002390 hypersensitivity pneumonitis
C0182792 hypersensitivity pneumonitis antibody determination re-
agents
```