

# Identifying Experimental Techniques in Biomedical Literature

**Meeta Oberoi, Craig A. Struble**

Dept. of Math., Stat., and Comp. Sci.

Marquette University

Milwaukee, WI 53201-1881

{meeta.oberoi, craig.struble}@marquette.edu

**Sonia Sugg**

Department of Surgery

Medical College of Wisconsin

Milwaukee, WI 53226

ssugg@mcw.edu

Named entity recognition of gene names, protein names, cell-lines, and other biologically relevant concepts has received significant attention by the research community. In this work, we considered named entity recognition of *experimental techniques* in biomedical articles. In our system to mine gene and disease associations, each association is categorized by the techniques used to derive the association. Categories are used to weight or remove associations, such as removing associations derived from microarray experiments.

We report on a pilot study to identify experimental techniques. Three main activities are discussed: *manual annotation*, *lexicon-based tagging*, and *document classification*. Analysis of manual annotation suggests several interesting linguistic characteristics arise. Two lexicon-based tagging approaches demonstrate little agreement, suggesting sophisticated tagging algorithms may be necessary. Document classification using abstracts and titles is compared with full-text classification. In most cases, abstracts and titles show comparable performance to full-text.

**Corpus** We built a corpus around our interest in gene associations with breast cancer to leverage the domain expertise of the authors. The corpus consisted of 247 sampled from 2571 papers associating breast cancer with a human gene in EntrezGene.

**Manual Annotation** Manual annotation was primarily performed by a graduate student in bioinformatics and a computer science Ph.D. with a research emphasis in bioinformatics. Annotators were instructed to highlight direct mentions of experimen-

tal techniques. During the study, we noted low inter-annotator agreement and stopped the manual process after annotating 102 of 247 documents.

Results were analyzed for linguistic characteristics. Experimental technique mentions appear with varying frequency in 6 typical document sections: *Title*, *Abstract*, *Introduction*, *Materials and Methods*, *Results* and *Discussion*. In some sections, such as *Introduction*, mentions are often for references and not the current document. Techniques such as *transfection* and *immunoblotting*, demonstrated diverse morphology. Other characteristics included use of synonyms and abbreviations, conjunctive phrases, and endophora.

**Tagging** Tagging is commonly used for named entity recognition. In our context, associations are categorized by generating a list of all tagged techniques tagged in a document.

Two taggers were tested on 247 documents to investigate the efficacy of two lexicons — MeSH and UMLS — containing experimental techniques. One approach used regular expressions for terms and permuted terms in the *Investigative Techniques* MeSH subhierarchy. The other used a natural language approach based on MetaMap Transfer (Aronson, 2001), mapping text to UMLS entries with the *Laboratory Procedure* semantic type (ID: T059).

Low inter-annotator agreement between taggers was exhibited with a maximum  $\kappa$  of 0.220. Both taggers exhibited limitations — failing to properly tag phrases such as “Northern and Western analyses” — and neither one is clearly superior to the other.

Technique	Full-Text		Abstract	
	1,000	All(59,628)	1,000	All(4,395)
Electrophoresis (144)	72.4/81.2/76.5	70.3/77.4/73.7	68.9/79.8/74.0	69.2/75.3/72.1
Western Blot Analysis (132)	71.3/83.6/77.0	71.4/77.0/74.1	67.5/79.8/73.1	68.2/76.2/72.0
Gene Transfer Technique (137)	76.3/92.1/83.4	74.6/88.6/81.0	77.0/89.6/82.8	76.2/88.3/81.8
Pedigree(10)	52.0/91.3/66.2	81.2/72.5/76.6	42.9/77.7/55.3	59.9/58.3/59.1
Sequence Alignment (24)	53.0/66.6/59.1	96.6/17.9/30.1	61.2/59.5/60.3	67.0/36.5/47.2
Statistics (107)	70.7/57.7/63.5	70.3/60.8/65.2	73.6/58.6/65.2	71.5/63.5/67.2

Table 1: Precision/Recall/F1-scores for classifiers with different vocabulary sizes.

**Document Classification** Document classification was also used to obtain a list of utilized experimental techniques. Each article is assigned to one or more classes corresponding to techniques used to generate results.

Two distinct questions were investigated. First, how well does classification perform if only the abstract and title of the article are available? Second, how does vocabulary size affect the classification?

Multinomial Naïve Bayes models were implemented in Rainbow (McCallum, 1996; McCallum and Nigam, 1998) for 24 MeSH experimental techniques. Document frequency in each class ranged from 10 (Pedigree) to 144 (Electrophoresis). Vocabularies consist of top information gain ranked words. Classifiers were evaluated by precision, recall, and F1-scores averaged over 100 runs. The corpus was split into 2/3 training and 1/3 testing, randomly chosen for each run.

Selected results are shown in Table 1. Full-text classifiers performed better than abstract based classifiers with a few exceptions: “Sequence Alignment” and “Gene Transfer Techniques”. The performance of abstract and full-text classifiers is comparable: F1 scores often differ by less than 5 points. Smaller vocabularies tend to improve the recall and overall F1 scores, while larger ones improved precision. Classifiers for low frequency (< 25) techniques generally performed poorly. One class, “Pedigree”, performed surprisingly well, with a maximum F1 of 76.6.

Considering that Naïve Bayes models are often baseline models and the small size of the corpus, classification performance is good.

**Related and Future Work** For comprehensive reviews of current work in biomedical literature mining, refer to (Cohen and Hersh, 2005) and (Krallinger et al., 2005). As future work, we will continue manual annotation, validate the informative capacity of sections with experiments similar to Sinclair and Webber (Sinclair and Webber, 2004), and investigate improvements in tagging and classification.

## References

- A. Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proc AMIA 2001*, pages 17–21.
- Aaron M. Cohen and William R. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6:57–71.
- Martin Krallinger, Ramon Alonso-Allende Erhardt, and Alfonso Valencia. 2005. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, 10:439–445.
- Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for Naïve Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*.
- Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/~mccallum/bow>.
- Gail Sinclair and Bonnie Webber. 2004. Classification from Full Text: A Comparison of Canonical Sections of Scientific Papers. In Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors, *COLING (NLPBA/BioNLP)*, pages 69–72, Geneva, Switzerland.