Exploring Two Biomedical Text Genres for Disease Recognition

Aurélie Névéol, Won Kim, W. John Wilbur, Zhiyong Lu*

National Center for Biotechnology Information
U.S. National Library of Medicine
Bethesda, MD 20894, USA
{neveola,wonkim,wilbur,luzh}@ncbi.nlm.nih.gov

Abstract

In the framework of contextual information retrieval in the biomedical domain, this paper reports on the automatic detection of disease concepts in two genres of biomedical text: sentences from the literature and PubMed user queries. A statistical model and a Natural Language Processing algorithm for disease recognition were applied on both corpora. While both methods show good performance (F=77% *vs.* F=76%) on the sentence corpus, results on the query corpus indicate that the statistical model is more robust (F=74% *vs.* F=70%).

1 Introduction

Contextual Information Retrieval (IR) is making use of additional information or assumptions about the users' needs beyond the obvious intent of the query. IR systems need to go beyond the task of providing generally relevant information by assisting users in finding information that is relevant to them and their specific needs at the time of the search. A practical example of a Google contextual IR feature is when the search engine returns a map showing restaurant locations to a user entering a query such as "Paris restaurants."

The *contextual* aspects of a user's search were defined for example by Saracevic (1997) who discussed integrating the cognitive, affective, and situational levels of human computer interaction in IR systems. Other research efforts studied users' search behavior based on their level of domain knowledge (Zhang et al., 2005) or aimed at modeling users' interests and search habits (Rose and Levinson, 2004; Teevan et al., 2005).

Information about the search context may be sought explicitly from the user through profiling or relevance feedback (Shen et al., 2005). Recent work also exploited query log analysis and basic computer environment information (Wen et al. 2004), which involve no explicit interaction with the user. In adaptive information retrieval, context information is inferred based on query analysis and collection characteristics (Bai and Nie 2008).

In the biomedical domain, a need for contextual information retrieval was identified in particular for clinical queries submitted to PubMed (Pratt and Wasserman, 2000). Building on the idea that a specific type of document is required for searches with a "clinical" context, the PubMed Clinical Queries portal was developed (Haynes and Wilczynski, 2004). A perhaps more prominent contextual feature of PubMed is the "citation sensor", which identifies queries classified by Rose and Levinson as reflecting a "Navigational" or "Obtain resource" goal. For example, the citation sensor will identify and retrieve a specific citation if the user enters the article title as the query. The analysis of Entrez logs shows that MEDLINE is the most popular database among the 30 or so databases maintained by the National Center for Biotechnology Information (NCBI) as it receives most of Entrez traffic. This suggests that there is a need to complement the information retrieved from MEDLINE by giving contextual access to other NCBI resources re-

Proceedings of the Workshop on BioNLP, pages 144–152,
Boulder, Colorado, June 2009.

levant to users' queries, such as Entrez Gene, Clinical Q&A or BookShelf. In addition, the NLM estimated that about 1/3 of PubMed users are not biomedical professionals. In this light, providing an access point to consumer information such as the Genetics Home Reference might also be useful. To achieve this, the *sensor* project was recently launched with the goal of recognizing a variety of biomedical concepts (e.g. gene, protein and drug names) in PubMed queries. These high-level concepts will help characterize users' search context in order to provide them with information related to their need beyond PubMed. For instance, if a user query contains the drug name "Lipitor", it will be recognized by the drug sensor and additional information on this drug from Clinical Q&A will be shown in the side bar in addition to default PubMed results. Since disease names are common in PubMed queries, the goal of this work is to investigate and benchmark computational techniques for automatic disease name recognition as an aid to implementing PubMed search contexts.

2 Related Work

Despite a significant body of literature in biomedical named entity recognition, most work has been focused on gene, protein, drug and chemical names through challenges such as BioCreAtIvE[1] or the TREC Genomics/Chemical tracks (Park and Kim, 2006). Other work addressed the identification of "medical problems" in clinical text (Aronson et al. 2007; Meystre and Haug, 2005). This task was the topic of a Medical NLP challenge[2], which released a corpus of anonymized radiography reports annotated with ICD9 codes. Although there is some interest in the biomedical community in the identification of disease names and more specifically the identification of relationships between diseases and genes or proteins (Rindflesh and Fizman, 2003), there are very few resources available to train or evaluate automatic disease recognition systems. To the best of our knowledge, the only publicly available corpus for disease identification in the literature was developed by Jimeno et al. (2008). The authors annotated 551 MEDLINE sentences with UMLS concepts and used this dataset to benchmark three different automatic methods for disease name recognition. A MEDLINE corpus annotated with "malignancy" mentions and part-of-speech tags is also available (Jin et al. 2006). This corpus is targeted to a very restricted type of diseases. The annotations are also domain specific, so that "cancer of the lung" is not considered a malignancy mention but a mention of malignancy and a mention of malignancy location.

As in previous studies, we aim to investigate the complexity of automatic disease recognition using state-of-the-art computational techniques. This work is novel in at least three aspects: first, in addition to using the MEDLINE sentence corpus (Jimeno et al 2008), we developed a new corpus comprising disease annotations on 500 randomly selected PubMed queries. This allowed us to investigate the influence of *local context*[3] through the comparison of system performance between two different genres of biomedical text. Second, by using a knowledge based tool previously benchmarked on the same MEDLINE corpus (Jimeno et al. 2008), we show that significant performance differences can be observed when parameters are adjusted. Finally, a state-of-the-art statistical approach was adapted for disease name recognition and evaluated on both corpora.

3 Two Biomedical Corpora with *disease* annotations

The first issue in the development of such a corpus is to define the very concept of disease. Among the numerous terminological resources available, such as Medical Subject Headings (MeSH®, 4,354 disease concepts) or the International Classification of Diseases (ICD9, ~18,000 disease concepts), the UMLS Metathesaurus® is the most comprehensive: the 2008AB release includes 252,284 concepts in the *disorder* Semantic Group defined by McCray et al. (2001). The UMLS Metathesaurus is part of the Semantic Network, which also includes a set of broad subject categories, or Semantic Types, that provide a consistent categorization of all concepts represented in the Metathesaurus. The Semantic Groups aim at providing an even broader categorization for UMLS concepts. For example, the disorder Semantic Group comprises 12 Semantic Types including *Disease or Syndrome*, *Cell or Molecular Dysfunction* and *Congenital Abnormalities*.

[1] http://biocreative.sourceforge.net/

[2] http://www.computationalmedicine.org/challenge/index.php

[3] Here, by *context*, we mean the information surrounding a disease mention available in the corpora. This is different from the "search context" previously discussed.

Furthermore, like the gene mention (Morgan et al. 2008) and gene normalization (Smith et al. 2008) tasks in BioCreative II, the task of disease name recognition can also be performed at two different levels:

1. *disease mention*: the detection of a snippet of text that refers to a disease concept (e.g. "alzheimer" in the sample query shown in Table 2)
2. *disease concept*: the recognition of a controlled vocabulary disease concept (e.g. "C0002395-alzheimer's disease" in our Table 2 example) in text.

In this work, we evaluate and report system performance at the concept level.

3.1 Biomedical literature corpus

Sentence	Kniest dysplasia is a moderately severe chondrodysplasia phenotype that results from mutations in the gene for type ii collagen col2a1.
Annotations	C0265279-Kniest dysplasia C0343284-Chondrodysplasia, unspecified

Table 1: Excerpt of literature corpus (PMID: 7874117)

The corpus made available by Jimeno et al. consists of 551 MEDLINE sentences annotated with UMLS concepts or concept clusters: concepts that were found to be linked to the same term. For example, the concepts "Pancreatic carcinoma" (C0235974) and "Malignant neoplasm of pancreas" (C0346647) share the same synonym "Pancreas Cancer", thus they were clustered. The sentences were selected from a set of articles curated for Online Mendelian Inheritance in Man (OMIM) and contain an average of 27(+/- 11) tokens, where tokens are defined as sequences of characters separated by white space. A set of UMLS concepts (or clusters) is associated with each sentence in the corpus. However, no boundary information linking a phrase in a sentence to an annotation was available. Table 1 shows a sample sentence and its annotations.

3.2 Biomedical query corpus

A total of 500 PubMed queries were randomly selected and divided into two batches of 300 and 200 queries, respectively. Queries were on average 3.45(+/- 2.64) tokens long in the 300 query batch and 3.58(+/- 4.63) for the 200 query batch, which is consistent with the average length of PubMed queries (3 tokens) reported by Herskovic et al. (2007).

The queries in the first set were annotated using Knowtator (Ogren, 2006) by three annotators with different backgrounds (one biologist, one information scientist, one computational linguist). Two annotators annotated the queries using UMLS concepts from the *disorder* group, while the other annotator simply annotated diseases without reference to UMLS concepts. Table 2 shows a sample query and its annotations. A consensus set was obtained after a meeting between the annotators where diverging annotations were discussed and annotators agreed on a final, unique, version of all annotations. The consensus set contains 89 disease concepts (76 unique).

Query	alzheimer csf amyloid
Annotations	Ann. 1: "alzheimer"; 0-8; Ann. 2, 3: "alzheimer"; 0-8; C0002395-alzheimer's disease

Table 2: Excerpt of annotated 300-query corpus. Boundary information is given as the character interval of the annotated string in the query (here, 0-8).

The queries in the second set were annotated with UMLS concepts from the *disorder* group by one of the annotators who also worked on the previous set. In this set, 53 disease concepts were annotated (51 unique).

4 Automatic *disease* recognition

With the perspective of a contextual IR application where the disease concepts found in queries will be used to refer users to disease-specific information in databases other than MEDLINE, we are concerned with high precision performance. For this reason, we decided to experiment with methods that showed the highest precision when compared to others. In addition, given the size of the corpora available and the type of the annota-

tions, machine learning methods such as CRFs or SVM did not seem applicable.

Table 3 shows a description of the training and test sets for each corpus.

Data	Lit. Corpus	Query Corpus
Training	276 sentences (487 disease concepts, 185 unique)	300 queries (89 disease concepts, 76 unique)
Testing	275 sentences (437 disease concepts, 185 unique)	200 queries (53 disease concepts, 51 unique)
All	551 sentences (924 disease concepts, 280 unique)	500 queries (142 disease concepts, 120 unique)

Table 3: Description of the training and test sets

4.1 Natural Language Processing

Disease recognition was performed using the Natural Language Processing algorithm implemented in MetaMap (Aronson, 2001)[4]. The tool was restricted to retrieve concepts from the *disorder* group, using the UMLS 2008AB release and "longest match" feature.

In practice, MetaMap parses the input text into noun phrases, generates variants of these phrases using knowledge sources such as the SPECIALIST lexicon, and maps the phrases to UMLS concepts.

4.2 Priority Model

The priority model was first introduced in (Tanabe and Wilbur, 2006) and is adapted here to detect *disease mentions* in free text. Because our evaluation is performed at the *concept* level, the mentions extracted by the model are then mapped to UMLS using MetaMap.

The priority model approach is based on two sets of phrases: one names of diseases, *D*, and one names of non-diseases, *N*. One trains the model to assign two numbers, *p* and *q*, to each token *t* that appears in a phrase in either *D* or *N*. Roughly, *p* is the probability that a phrase from *D* or *N* that has the token *t* in it is actually from *D* and *q* is the relative weight that should be assigned to *t* for this purpose and represents a quality estimate. Given a phrase

$$ph = t_1 t_2 \ldots t_k \quad (1)$$

and for each t_i the corresponding numbers p_i and q_i we estimate the probability that $ph \in D$ by

$$prob = p_1 \prod_{j=2}^{k} 1-q_j \; + \sum_{i=2}^{k} q_i p_i \prod_{j=i+1}^{k} 1-q_j \quad (2)$$

The training procedure for the model actually chooses the values of all the p and q quantities to optimize the $prob$ values over all of *D* and *N*. For this work we have extended the approach to include a quantity

$$qual = \left[q_1 p_1 \prod_{j=2}^{k} 1-q_j \; + \sum_{i=2}^{k} q_i^2 p_i \prod_{j=i+1}^{k} 1-q_j \right] \Big/ prob \quad (3)$$

which represents a weighted average of all the quality numbers q_i. We apply this formula to obtain $qual$ as long as $prob \geq 0.5$. If $prob < 0.5$ we replace all numbers p_i by $1-p_i$ in (2) and (3) to obtain $qual$.

For this application we obtained the sets *D* and *N* from the SEMCAT data (Tanabe, Thom et al. 2006) supplemented with the latest UMLS data. We removed any term from *D* and *N* that contained less than five characters in order to decrease the occurrence of ambiguous terms. Also the 1,000 most frequent terms from *D* were examined manually and the ambiguous ones were removed. The end result is a set of 332,984 phrases in *D* and 4,253,758 phrases in *N*. We trained the priority model on *D* and *N* and applied the resulting training to compute for each phrase in *D* and *N* a vector of values $prob, qual$. In this way *D* and *N* are converted to V_D and V_N. We then constructed a Mahalanobis classifier (Duda, Hart and Stork, 2001) for two dimensional vectors as the difference in the Mahalanobis distance of any such vector to Gaussian approximations to V_D and V_N. We refer to the number produced by this classifier as the Mahalanobis score. By randomly dividing both *D* and *N* into three equal size pieces and training on two from each and testing on the third, in a three-fold cross validation we found the Mahalanobis classifier to perform at 98.4% average precision and 93.9% precision-recall breakeven point. In a final step we applied a simple regression method to estimate the probability that a given Maha-

[4] Additional information is also available at http://metamap.nlm.nih.gov/

lanobis score was produced by a phrase belonging to *D* and not *N*. Given a phrase *phr* we will denote this final probability produced as PMA(*phr*).

The second important ingredient of our statistical process is how we produce phrases from a piece of text. Given a string of text *TX* we apply tokenization to *TX* to produce an ordered set of tokens $t_1, t_2, \ldots, t_n$. Among the tokens produced will be punctuation marks and stop words and we denote the set of all such tokens by Z . We call a token segment $t_j, \ldots, t_k$ maximal if it contains no element of Z and if either $j=1$ or $t_{j-1} \in Z$ and likewise if $k=n$ or $t_{k+1} \in Z$. Given text *TX* we will denote the set of all maximal token segments produced in this way by $S_{\max}(TX)$. Now given a maximal token segment *mts*= $t_j, \ldots, t_k$ we define two different methods of finding phrases in *mts*. The first assumes we are given an arbitrary set of phrases *PH*. We recursively define a set of phrases $I\ mts, PH$ beginning with this set empty and with the parameter $u = j$. Each iteration consists of asking for the largest $v \le k$ for which $t_u, \ldots, t_v \in PH$. If there is such a v we add $t_u, \ldots, t_v$ to $I\ mts, PH$ and set $u = v+1$. Otherwise we set $u = u+1$. We repeat this process as long as $u \le k$. The second approach assumes we are given an arbitrary set of two token phrases *P2*. Again we recursively define a set of phrases $J\ mts, P2$ beginning with this set empty and with the parameter $u = j$. Each iteration consists of asking for the largest $v \le k$ for which given any $i,\ u \le i < v,\ \ t_i, t_{i+1} \in P2$. If there is such a v we add $t_u, \ldots, t_v$ to $J\ mts, P2$ and set $u = v+1$. Otherwise we set $u = u+1$. We repeat this process as long as $u \le k$.

In order to apply our phrase extraction procedures we need good sets of phrases. In addition to *D* and *N* already defined above, we use another set of phrases defined as follows. Let *R* denote the set of all token strings with two or more tokens which do not contain tokens from Z and for which there are at least three MEDLINE records (title and abstract text only) in which the token string is repeated at least twice.

We then define $R' = R - D \cup N$. We make use of R' in addition to *D* and *N*. For the set $P2$ we take the set of all two token phrases in MEDLINE documents for which the two tokens co-occur as this phrase much more than expected, i.e., with a $\chi^2 \ge 10,000$ (based on the two-by-two contingency table).

#Initialization: Given a text *TX*, set $S \leftarrow S_{\max}\ TX$ and $X \leftarrow \varnothing$.
#Processing: While($S \neq \varnothing$){
I. select $mts \in S$
II. If($I\ mts, D \neq \varnothing$) $K \leftarrow I\ mts, D$
else if($I\ mts, R' \neq \varnothing$) $K \leftarrow I\ mts, R'$
else if($I\ mts, N \neq \varnothing$) $K \leftarrow \varnothing$
else
if($J\ mts, P2 \neq \varnothing$) $K \leftarrow J\ mts, P2$
else $K \leftarrow \varnothing$
III. $X \leftarrow X \cup K$
IV. $S \leftarrow S - mts$
}
#Return: All pairs $phr, PMA\ phr\ ,\ phr \in X$

Figure 1: Phrase finding algorithm

With these preliminaries, our phrase finding algorithm in pseudo-code is shown in Figure 1.

The output of this algorithm may then be filtered by setting a threshold on the *PMA* values to accept.

5 Results

5.1 Assessing the difficulty of the task

To assess the difficulty of disease recognition, we computed the inter-annotator agreement (IAA) on the 300-query corpus. Agreement was computed at the *disease mention* level for all three annotators and at the *disease concept* level for the two annotators who produced UMLS annotations.

Inter-annotator agreement measures for NLP applications have been recently discussed by Artstein and Poesio (2008) who advocate for the use of chance corrected measures. However, in our case, agreement was partly computed on a very large set of categories (UMLS concepts) so we decided to use Knowtator's built-in feature, which computes IAA as the percentage of agreement and

allows partial string matches. For example, in the query "dog model transient ischemic attacks", annotator 1 selected "ischemic attacks" as a *disorder* while annotator 2 and 3 selected "transient ischemic attacks" as UMLS concept *C0007787: Attacks, Transient Ischemic*. In this case, at the subclass level ("disorder") we have a match for this annotation. But at the exact span or exact category level, there is no match. Table 4 shows details of IAA at the disease mention level when partial matches are taken into account. For exact span matches, the IAA is lower, at 64.87% on average.

Disorder IAA	**Ann. 1**	**Ann. 2**	**Ann. 3**
Ann. 1	100%	71.77%	75.86%
Ann. 2		100%	71.68%
Ann. 3			100%

Table 4: Agreement on disease mention annotations (partial match allowed) – **average is 73.10%**

At the concept level, the agreement (when partial matches were allowed) varied significantly depending on the semantic types. It ranged between 33% for *Findings* and 83% for *Mental or Behavioral Dysfunction*. However, agreement on the most frequent category, *Disease or Syndrome*, was 72%, which is close to the annotators' overall agreement at the mention level. One major cause of disagreement was ambiguity caused by concepts that were clustered by Jimeno et al. For example, in query "osteoporosis and "fracture pattern", annotator 2 marked "osteoporosis" with both "C0029456-osteoporosis"(a *Disease or Syndrome* concept) and "C1962963-osteoporosis adverse event"(a *Finding* concept) while annotator 3 only used "C0029456-osteoporosis".

5.2 Results on Literature corpus

As shown in Table 3, the corpus was randomly split into a training set (276 sentences) and a test set (275 sentences). The training set was used to determine the optimal probability threshold for the Priority Model and parameter selection for MetaMap, respectively.

Priority Model parameter adjustments: the first result observed from applying the Priority Model was that *D* yielded about 90% of the output of the algorithm. Also results coming from R' and $P2$ were not well mapped to UMLS concepts by MetaMap. As a result, in this work we ignored disease candidates retrieved based on R' and $P2$. The best F-measure was obtained for a threshold of 0.3, which was consequently used on the test set.

Since the Priority Model algorithm does not perform any mapping to a controlled vocabulary source, the mapping was performed by applying MetaMap to the snippets of text returned with a probability value above the threshold.

Threshold	**P**	**R**	**F**
0	64	73	67
.1	67	73	70
.2	67	73	70
.3	**68**	**73**	**71**
.4	68	73	70
.5	68	72	69
.6	68	72	69
.7	68	72	69
.8	68	68	68
.9	65	60	62

Table 5: Precision (P), Recall (R) and F-measure of the Priority Model on the training set for different values of the probability threshold.

The results presented in Table 5 were obtained before any MetaMap adjustments were made.

MetaMap parameter adjustments: an error analysis was performed to adjust MetaMap settings. Errors fell into the following categories:

- A more specific disease should have been recognized (e.g. "deficiency" *vs.* "C2 deficiency")
- The definition of a cluster was lacking (e.g. "G6PD deficiency" comprised C0237987- Glucose-6-phosphate dehydrogenase deficiency anemia and C0017758- Glucosphosphate Dehydrogenase Deficiency but not C0017920- Deficiency of glucose-6-phosphatase)
- MetaMap mapping was erroneous (e.g. "hereditary breast" was mapped to C0729233-Dissecting aneurysm of the thoracic aorta instead of C0346153-Hereditary Breast Cancer)

The results of inter-annotator agreement and further study of MetaMap mappings indicated that concepts with the semantic type *Findings* seemed

to be frequently retrieved erroneously. For this reason, we also experimented not taking *Findings* into account as an additional adjustment for MetaMap. Table 6 shows the results of applying the MetaMap adjustments yielded from the error analysis on the training corpus.

Threshold	Findings	P	R	F
.3	Yes	80	78	79
.3	**No**	**85**	**78**	**81**

Table 6: performance of the Priority Model on the training set for threshold .3 depending on whether mappings to *Findings* are used in the "adjustments"

MetaMap disorder detection was also performed directly on the training corpus. An error analysis similar to what was presented above was carried out to determine the best parameters. Table 7 below shows the results obtained when all concepts from the 12 Semantic Types (STs) in the *disorder* group are taken into account with no adjustments ("raw"). Then, results including the adjustments from the error analysis are shown when all 12 STs are taken into account, when *Findings* are excluded (11STs) and when only the most frequent 6STs in the training set are taken into account.

Processing	P	R	F
Raw (12 STs)	50	77	61
Adjusted (12 STs)	52	75	61
Adjusted (11 STs)	57	73	64
Adjusted (6 STs)	**77**	**72**	**74**

Table 7: Performance of MetaMap on the training set

Finally, Table 8 shows the performance of both methods on the test set, using the optimal settings determined on the training set:

Method	P	R	F
Priority Model	**80**	**74**	**77**
MetaMap	75	78	76

Table 8: Precision (P), Recall (R) and F-measure of the Priority Model and MetaMap on the test set

5.3 Results on Query Corpus

The 300-query corpus was used as a training set and the 200-query corpus was used as a test set. For consistency with work on the literature corpus, we assessed the disease recognition on a gold standard set including "clusters" of UMLS concepts were appropriate. As previously with the Literature corpus, we used the training set to determine the best settings for each method. The performance of the Priority Model at different values of the probability threshold, based on the use of *D* and *N* as the sets of sample phrases is similar to that obtained with the literature corpus; 0.3 stands out as one of the three values for which the best F-measure is obtained (tied with .5 and .8).

Because of the brevity of queries *vs.* sentences, the MetaMap error analysis was very succinct and resulted in:

- Removal of C0011860-Diabetes mellitus type 2 as mapping for "diabetes"
- Removal of all occurrences of C0600688-Toxicity and C0424653-Weight symptom (finding)
- Adjustment on the number of STs taken into account

The difference in performance obtained on the training set for the different MetaMap adjustments considered is shown in Table 9 when MetaMap was applied to Priority Model output and in Table 10 when it was applied directly on the queries.

Threshold	Findings	P	R	F
.3	Yes	60	72	65
.3	**No**	**73**	**70**	**71**

Table 9: performance of the Priority Model on the training set for threshold .3 depending on whether mappings to *Findings* are used in the "adjustments"

Processing	P	R	F
Raw (12 STs)	41	82	55
Adjusted (12 STs)	44	82	57
Adjusted (11 STs)	58	81	68
Adjusted (6 STs)	**64**	**75**	**69**

Table 10: performance of MetaMap on the training set

Finally, Table 11 shows the performance of both methods on the test set, using the optimal settings determined on the training set:

Method	P	R	F
Priority Model	**76**	**72**	**74**
MetaMap	66	74	70

Table 11: Precision (P), Recall (R) and F-measure of the Priority Model and MetaMap on the test set

6 Discussion

Comparing the Two Methods. The performance of both methods on the query corpus is comparable to inter-annotator agreement (F=70-74 *vs.* IAA=72 on *Disease and Syndromes*). On both corpora, the Priority Model achieves higher precision and F-measure, while MetaMap achieves better recall.

Comparing the results obtained with MetaMap with those reported by Jimeno et al., precision is lower, but recall is much higher. This is likely to be due to the different MetaMap settings, and the use of different UMLS versions - Jimeno et al. did not provide any of this information, but based on the publication date of their paper, it is likely that they used one of the 2006 UMLS releases. Meystre and Haug (2006) also found that significant performance differences could be obtained with MetaMap by adjusting the content of the knowledge sources used.

On both text genres, 0.3 was found to be the optimal probability threshold for the Priority Model. Based on the performance at different values of the threshold, it seems that the model is quite efficient at ruling out highly unlikely diseases. However, for values above .3 the performance does not vary greatly.

Comparing Text Genres. For both methods, disease recognition seems more efficient on sentences. This is to be expected: sentences provide more context (e.g. more tokens surrounding the disease mention are available) and allow for more efficient disambiguation, for example on acronyms. Although acronyms are frequent both in queries and sentences, more undefined acronyms are found in queries. However, the difference in performance between the two methods seems higher on the query corpus. This indicates that the Priority Model could be more robust to sparse context.

It should be noted that there were diseases in all sentences in the literature corpus *vs.* about 1/3 to 1/2 of the queries. In addition, the query corpus included many author names, which could create confusion with disease names (in particular for the Priority Model). This difficulty was not found in the sentence corpus. However, sentences sometimes contain negated mention of diseases, which never occurred in the query corpus where little to no syntax is used.

We also noticed that while *Findings* seemed to be generally problematic concepts in both corpora, other concepts such as *Injury and Poisoning* were much more prevalent in the query corpus. For this reason, for the general task of disease recognition, a drastic restriction to as little as 6 STs is probably not advisable.

Limitations of the study. One limitation of our study is the relatively small number of disease concepts in the query corpus. Although the query and sentence corpus contain about 500 queries/sentences each, there are significantly less disease concepts found in queries compared to sentences. As a result, there is also less repetition in the disease concept found. This is partly due to the brevity of queries compared to sentences but mainly to the fact that while all the sentences in the literature corpus had at least one disease concept, this was not the case for the query corpus. We are currently addressing this issue with the ongoing development of a large scale query corpus annotated for diseases and other relevant biomedical entities.

7 Conclusions

We found that of the two steps of disease recognition, disease mention gets the higher inter-annotator agreement (*vs.* concept mapping). We have applied a statistical and an NLP method for the automatic recognition of disease concepts in two genres of biomedical text. While both methods show good performance (F=77% vs. F=76%) on the sentence corpus, results indicate that the statistical model is more robust on the query corpus where very little disease context information is available (F=74% *vs.* F=70%). As a result, the priority model will be used for disease detection in PubMed queries in order to characterize users' search contexts for contextual IR.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would like to thank S. Shooshan and T. Tao for their contribution to the annotation of the query corpus; colleagues in the NCBI engineering branch for their valuable feedback at every step of the project.

References

Alan R. Aronson, Olivier Bodenreider, Dina Demner-Fushman, Kin Wah Fung, Vivan E. Lee, James G. Mork et al. 2007. From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. ACL Workshop BioNLP.

Alan Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of AMIA Symp*:17-21.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4): 555-596

Jing Bai, and Jian-Yun Nie. 2008. Adapting information retrieval to query contexts. *Information Processing & Management.* 44(6):1902-22

Robert O. Duda, Peter. E. Hart and David G. Stork. 2001. *Pattern Classification.* New York: John Wiley & Sons, Inc.

R. Brian Haynes and Nancy L. Wilczynski. 2004. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ*. 328(7447):1040.

Jorge R. Herskovic, Len Y. Tanaka, William Hersh and Elmer V. Bernstam. 2007. A day in the life of PubMed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association.* 14(2):212-20.

Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga and Dietrich Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*. 11;9 Suppl 3:S3.

Yang Jin, Ryan T McDonald, Kevin Lerman, Mark A Mandel, Steven Carroll, Mark Y Liberman et al. 2006. Automated recognition of malignancy mentions in biomedical literature. *BMC Bioinformatics.* 7:492.

Alexa T. McCray, Anita Burgun and Olivier Bodenreider. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Proceedings of Medinfo* 10(Pt 1):216-20.

Stéphane Meystre and Peter J. Haug. 2006. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform.* 39(6):589-99.

Alexander A. Morgan, Zhiyong Lu, Xinglong Wang, Aaron M. Cohen, Juliane Fluck, Patrick Ruch et al. 2008. Overview of BioCreative II gene normalization. *Genome Biol.* 9 Suppl 2:S3.

Phillip V. Ogren. 2006. Knowtator: A plug-in for creating training and evaluation data sets for Biomedical Natural Language systems. *9th Intl. Protégé Conference*

Jong C. Park and Jung-Jae Kim. 2006. Named Entity Recognition. In S. Ananiadou and J. McNaught (Eds), *Text Mining for Biology and Biomedicine* (pp. 121-42). Boston|London:Artech House Inc.

Wanda Pratt and Henry Wasserman. 2000. QueryCat: automatic categorization of MEDLINE queries. *Proceedings of AMIA Symp*:655-9.

Tom C. Rindflesh and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform.* 36(6):462-77

Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international Conference on World Wide Web*:13-9

Tefko Saracevic. 1997. The Stratified Model of Information Retrieval Interaction: Extension and Application. *Proceedings of the 60th meeting of the. American Society for Information Science*:313-27

Xuehua Shen, Bin Tan and ChengXiang Zhai. 2005 Context-sensitive information retrieval using implicit feedback, In *Proceedings of the 28th annual international conference ACM SIGIR conference on Research and development in information retrieval*: 43-50.

Larry Smith, Laurraine K. Tanabe, Rie J. Ando, Cheng-Ju Kuo, I-Fang Chung , Chun-Nan Hsu et al. 2008. Overview of BioCreative II gene mention recognition. Genome Biol. 9 Suppl 2:S2.

Laurraine K. Tanabe, Lynn. H. Thom, Wayne Matten, Donald C. Comeau and W. John Wilbur. 2006. SemCat: semantically categorized entities for genomics. *Proceedings of AMIA Symp*: 754-8.

Laurraine K. Tanabe and W. John Wilbur. 2006. A Priority Model for Named Entities. *Proceedings of HLT-NAACL BioNLP Workshop*:33-40

Jaime Teevan, Susan T. Dumais and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Proceeding of ACM-SIGIR'05*:449–56.

Ji-Rong Wen, Ni Lao, Wei-Ying Ma. 2004. Probabilistic model for contextual retrieval. *Proceedings of ACM-SIGIR'04*:57–63

Xiangmin Zhang, Hermina G.B. Anghelescu and Xiaojun Yuan. 2005. Domain knowledge, search behavior, and search effectiveness of engineering and science students: An exploratory study, *Information Research* 10(2): 217.