

ISMB BioLINK SIG Linking Literature,
Information and Knowledge for Biology

Special Interest Group on Text Data Mining
In association with ISMB 2004, Glasgow, Scotland
July 29, 2004

The BioLINK Text Data Mining SIG will take place from 9:00- 17:00 on July 29, 2004
(<http://www.iscb.org/ismbeccb2004/>)

This year, the BioLINK SIG meeting will focus on resources and tools for text mining, with special emphasis on the evaluation of these tools. We accepted 5 of 11 submissions for presentation at the workshop. The paper topics cover a comparison of two evaluations, interannotator agreement, ontology-based extraction and two papers on lexical resources for text mining in biology.

We have invited speakers in the following areas:

- The recent BioCreative evaluation (Critical Assessment of Information Extraction in Biology) (<http://www.pdg.cnb.uam.es/BioLink/BioCreative.eval.html>)
- TREC Genomics track
<http://medir.ohsu.edu/~genomics/>
- The BioMINT project
<http://www.biomint.org/>
- CASP: Critical Assessment of Techniques for Protein Structure Prediction
<http://predictioncenter.llnl.gov/casp5/Casp5.html>
- EVA: Evaluation of automatic structure prediction servers
<http://cubic.bioc.columbia.edu/eva/cafasp/>

The meeting will include a final session to discuss future directions for the BioLINK SIG, especially related to the topic of critical assessment and evaluation.

Lynette Hirschman

Program Committee

Christian Blaschke	CNB Madrid
Luc Dehaspe	PharmaDM
Robert Gaizauskas	University of Sheffield
William Hersh	Oregon Health & Science University
Lynette Hirschman	MITRE
Alfonso Valencia	CNB Madrid
Karin Verspoor	Los Alamos National Laboratory
Alexander Yeh	MITRE

BioLINK SIG: Text Mining in Biology

29 July 2004, Glasgow, UK

Start	End	TITLE	Speaker/Authors
9:00	10:15	Report on BioCreAtIvE: Critical Assessment of Information Extraction in Biology	
		Task1: Gene Name Extraction and Normalization	Alexander Yeh, Lynette Hirschman
		Task 2: Functional Annotation	Christian Blaschke, Alfonso Valencia
10:15	10:45	Report on TREC Genomics Track	William Hersh
10:45	11:10	BREAK	
11:10	11:35	A System for Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations	Shipra Dingare, Malvina Nissim, Jenny Finkel, Claire Grover, Christopher Manning
11:35	12:00	Protein Name Tagging Guidelines: Lessons Learned	Inderjeet Mani, Zhangzhi Hu, Cathy Wu, Seok Bae Jang, Ken Samuel, Matthew Krause, Jon Phillips
12:00	1:00	LUNCH	
1:00	1:30	BioMinT: a database curator's assistant for biomedical text processing	Anne-Lise Veuthey
1:30	1:55	Ontology-Based Interactive Information Extraction from Scientific Abstracts	David Milward, Marcus Bjärelund, William Hayes, Michelle Maxwell, Lisa Öberg, Nick Tilford, James Thomas, Roger Hale, Sylvia Knight, Julie Barnes
1:55	2:20	A Web Service for Biomedical Term Look-Up	Rob Gaizauskas, Mark Hepple, Henk Harkema, Ian Robert, Neil Davis
2:20	2:45	Towards a Semantic Lexicon for Biological Language Processing	Karin Verspoor
2:45	3:15	BREAK	
3:15	4:00	Report on CASP and EVA	Anna Tramontano, Burkhard Rost
4:00	4:15	Announcements	
4:15	5:00	Discussion: Future directions for BioLINK SIG	Larry Hunter

yellow = invited speakers; green = submitted talks

A System For Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations

Shipra Dingare,* Jenny Finkel,** Malvina Nissim,*
Christopher Manning,** Claire Grover*

*Institute for Communicating and Collaborative Systems
{sdingar1|mnissim|grover}@inf.ed.ac.uk
University of Edinburgh, United Kingdom

**Department of Computer Science
{jrfinkel|manning}@cs.stanford.edu
Stanford University, United States

Abstract

We present a maximum-entropy based system for identifying Named Entities (NEs) in biomedical abstracts and present its performance in the only two biomedical Named Entity Recognition (NER) comparative evaluations that have been held to date, namely BioCreative and Coling BioNLP. Our system obtained an exact match f-score of 83.2% in the BioCreative evaluation and 70.1% in the BioNLP evaluation. We discuss our system in detail including its rich use of local features, attention to correct boundary identification, innovative use of external knowledge resources including parsing and web searches, and rapid adaptation to new NE sets. We also discuss in depth problems with data annotation in the evaluations which caused the final performance to be lower than the optimal.

1. Introduction

The explosion of information in the biomedical domain has led to immense interest in automated information extraction techniques and consequently to a number of publications describing systems and results for natural language processing tasks on biomedical data. With each group addressing varying tasks, using varying evaluation corpora, and employing varying scoring methods, it has been impossible to properly compare systems and assess the state of progress in the field. The use of standardized evaluations to remedy this state of affairs is only beginning; the Text Retrieval Conference only recently initiated a genomics track to assess biomedical information retrieval and question-answering. Here we focus on the task of Named Entity Recognition (NER) which requires identification of names in shallow semantic categories such as protein names or drug names. A number of groups have reported results on biomedical NER, attempting to identify anywhere between four and twenty-four categories, evaluating on corpora ranging from 30 to 100 abstracts and reporting scores varying from 3% for the class “RNA” to 92% for the specific protein “SH3” (Collier et al., 2000; Fukuda, 1998; Kazama et al., 2002; Nobata et al., 1999). Recently, two comparative evaluations have been held to evaluate the state of progress in the field: BioCreative (Blaschke et al., 2004) and Coling BioNLP (Collier et al., 2004).

In this paper we present a maximum-entropy based system for NER in biomedical abstracts which was entered in both of the above evaluations. Our system was originally designed for the BioCreative evaluation and was then adapted for the BioNLP task. We describe our system in detail including its exhaustive use of local context as well as exploitation of a variety of external resources including parsing, Google web-querying, and gazetteers. We present our results in both evaluations and consider how the quality

of the data affected the results. We found that performance in the tasks was more reflective of data quality than task difficulty. We discuss ways of improving annotation to provide maximal performance for machine learning systems.

2. The Tasks

The BioCreative NER task required participants to identify a single entity “NEWGENE” in biomedical abstracts. This entity corresponded roughly to gene and protein names. Organizers provided 10,000 sentences from MEDLINE abstracts as training data and 5000 sentences as evaluation data. The average number of entities per sentence was roughly similar in both training and evaluation data (approximately 1.19).

The BioNLP NER task required participants to identify the five NEs “protein”, “DNA”, “RNA”, “cell line” and “cell type” in medical abstracts. The task was based on the GENIA corpus (Ohta et al., 2002), a corpus of MEDLINE abstracts annotated for approximately 35 NE classes involved in biological reactions relating to transcription factors in human blood cells. The original set of NEs was collapsed into the above 5 by merging specific classes such as “protein molecule”, “protein family or group”, and “protein substructure” into broader classes (“protein”) and dropping other classes such as “body part” and “virus” completely; the nested annotations contained in the original corpus were also removed for simplicity. The organizers did not say whether the adaptation of the corpus for the BioNLP task was done automatically. The entire GENIA corpus of 18,546 sentences was provided as training data, and an additional 3,856 sentences as evaluation data. The average number of NEs per sentence was quite different between the training and evaluation data (for protein 1.63 in training versus 1.34 in testing, for DNA 0.51 vs 0.27, for RNA 0.05 vs 0.03, for cell line 0.20 vs 0.12, for cell type 0.36 vs 0.49).

Both BioNLP and BioCreative used the same exact-match scoring criterion in which participants were penalized twice, both as a false positive (FP) and as a false negative (FN), for an answer with incorrect boundaries. For example, if the correct entity was *human interleukin-2 gene* and the system returned only *interleukin-2*, the former would be counted as a FN and the latter as a FP.

3. System Description

Our system is a Maximum Entropy Markov Model (McCallum et al., 2000) with a Limited Memory Quasi-Newton maximizer based on a system used for the CoNLL 2003 shared task (Klein et al., 2003). The system essentially uses a logistic regression model to classify each word, overlaid with a Viterbi-style algorithm to find the best sequence of classifications. Maximum entropy models have been used with much success in NER tasks and are known for their ability to incorporate a large number of overlapping features. For both evaluations we devoted most of our efforts to finding useful features for the NERs required. The final system makes exhaustive use of clues within the sentence including character substrings, words, word shapes, and detection of abbreviations, as well as using longer-distance information obtained from the surrounding abstract and relations obtained by parsing, and various external resources, including a Google web-querying technique, the TnT part-of-speech tagger (Brants, 2000), and a gazetteer. We normalized names of months and days of the week to lowercase, and mapped the British spellings of a few common medical terms to their American equivalents. In the following sections we describe our full feature set.

We outline first the features utilizing the local context and secondly the features corresponding to external resources and larger context. We also describe a postprocessing phase aimed at reducing boundary errors. Our final systems for both evaluations employed over 1.25 million features.

3.1. Local Features

We used a variety of features describing the immediate context of each word, including the word itself, the previous and next words, bi-grams of the current word and next word and the current word and previous word, character n-grams up to a length of 6, word shapes, and features describing the named entity tags assigned to the previous words. Word shapes refer to mappings of words to simplified representations that encode attributes such as length and whether the word contains capitalization, numerals, greek letters, and so on. We also incorporated POS tags from the TnT tagger trained on the GENIA gold standard for POS in biomedical text. We made use of abbreviation matching to ensure consistency of labels between an abbreviation and its long form. A list of abbreviations and long forms was extracted from the data using the method of (Schwartz and Hearst, 2003); then all occurrences of the short and long forms in the data were labeled as such. (For BioNLP, we combined the list with the short and long forms from the BioCreative data.) Features referencing these labels were then included in the classifier. Following (Kazama et al., 2002) we added disjunctive word features. Lastly, a parentheses-matching

Word Features	w_i, w_{i-1}, w_{i+1}
	Last "real" word (BioCreat. only)
	Next "real" word (BioCreat. only)
	Disj. of 4 prev words (BioNLP - 5)
	Disj. of 4 next words (BioNLP - 5)
Bigrams	$w_i + w_{i-1}$
	$w_i + w_{i+1}$
TnT POS	$POS_i, POS_{i-1}, POS_{i+1}$
Character Substrings	Up to a length of 6 (BioNLP - prefix/suffix only)
Abbreviations	$abbr_i$
	$abbr_{i-1} + abbr_i$
	$abbr_i + abbr_{i+1}$
	$abbr_{i-1} + abbr_i + abbr_{i+1}$
Word Shape	$shape_i, shape_{i-1}, shape_{i+1}$
	$shape_{i-1} + shape_i$
	$shape_i + shape_{i+1}$
	$shape_{i-1} + shape_i + shape_{i+1}$
TnT POS + Word	$w_i + POS_i$
	$w_{i-1} + POS_i$
	$w_{i+1} + POS_i$
Word Shape + Word	$w_{i-1} + shape_i$
	$w_{i+1} + shape_i$
Shape + Word Disj (BioNLP only)	$shape_i + \text{Disj of 5 Prev Words}$
	$shape_i + \text{Disj of 5 Next Words}$
Previous NE	NE_{i-1}
	$NE_{i-2} + NE_{i-1}$
	$NE_{i-3} + NE_{i-2} + NE_{i-1}$ (BioNLP only)
	$NE_{i-4} + NE_{i-3} + NE_{i-2} + NE_{i-1}$ (BioNLP only)
Previous NE + Word	$NE_{i-1} + w_i$
Previous NE + POS	$NE_{i-1} + POS_{i-1} + POS_i$
	$NE_{i-2} + NE_{i-1} + POS_{i-2} + POS_{i-1} + POS_i$
	$NE_{i-3} + NE_{i-2} + NE_{i-1} + POS_{i-3} + POS_{i-2} + POS_{i-1} + POS_i$ (BioNLP only)
Previous NE + Abbr	$NE_{i-1} + abbr_{i-1} + abbr_i$
	$NE_{i-2} + NE_{i-1} + abbr_{i-2} + abbr_{i-1} + abbr_i$
Previous NE + Shape	$NE_{i-1} + shape_i$
	$NE_{i-1} + shape_{i+1}$
	$NE_{i-1} + shape_{i-1} + shape_i$
	$NE_{i-2} + NE_{i-1} + shape_{i-2} + shape_{i-1} + shape_i$
PrevNE+Shape+POS (BioNLP only)	$NE_{i-2} + NE_{i-1} + POS_{i-2} + POS_{i-1} + POS_i + shape_i$
	$NE_{i-3} + NE_{i-2} + NE_{i-1} + POS_{i-3} + POS_{i-2} + POS_{i-1} + POS_i + shape_i$
Paren-Matching	A feature that signals when one parentheses in a pair has been assigned a different tag than the other in a window of 4 words

Table 1: Local Features

feature that signalled when one parenthesis was classified differently from its pair was added in an effort to eliminate errors where the tagger classified matching parentheses differently. We combined all of the above base-level features in various ways. The full set of local features is outlined in Table 1.

3.2. External Resources and Larger Context

The features described here comprise various external resources including gazetteers, a web querying technique and relations obtained by parsing. The basic assumption behind and motivation for using external resources is that there are instances in the data where contextual clues do not provide sufficient evidence for confident classification. In such cases external resources may bridge the gap, either in the form of word lists known to refer to genes (gazetteers) or through examination of other contexts in which the same token appears and the exploitation of more indicative contexts (as with web-querying and use of surrounding text such as abstracts).

3.2.1. Deep Syntax Features

Our system benefits from relational information obtained by parsing. While it has been stated that full parsing of biomedical text is beyond current technology, we were able to successfully parse the BioNLP training and evaluation corpora using the Stanford Parser (Klein and Manning, 2003) operating on the TnT POS tags. Since we did not have parsed biomedical text with which to train the parser, we used the parsed Wall Street Journal; we believe that the unlexicalized nature of the Stanford parser made it suitable for parsing data from a different domain. For each word that appeared in a noun phrase, the head and governor of the noun phrase were extracted. These features were not useful in BioCreative because it involved identification of only one entity, but they were useful for BioNLP where one had to disambiguate between similar classes; (Shen et al., 2003) and (Nobata et al., 1999) also benefit from use of head noun features with the GENIA entities. This disambiguation requires longer distance information and a better representation of the context in which the word appears. For instance, the word *phosphorylation* occurs in the training corpus 492 times, 482 of which it was classified as other. However, it was the governor of 738 words, of which 443 were protein, 292 were other and only 3 were cell line.

3.2.2. Abstract

A number of NER systems have made effective use of how the same token was tagged in different parts of the same document (Mikheev et al., 1999; Curran and Clark, 2003). A token which appears in an unindicative context in one sentence may appear in a very obvious context in another sentence in the same abstract. To leverage this we tagged each abstract twice, providing for each token a feature indicating whether it was tagged as an entity elsewhere in the abstract. For BioCreative we were provided only single sentences from abstracts; we used cgi scripts to automatically obtain the corresponding full abstracts from MEDLINE. In a practical application this would be unnecessary since one would always have the full abstract. Abstract information was only useful when combined with information on frequency.

3.2.3. Web

As the largest corpus in existence, the web has been used effectively in a variety of NLP tasks (Keller and Lapata, 2003; Grefenstette, 1999; Markert et al., 2003). In our

use of the web we built several contexts indicative of target entities, such as “X gene” or “X antagonist” for genes, “X mRNA” for RNA, or “X ligation” for proteins. We then substituted the variable “X” with potential entities and submitted the resulting patterns to the web. We used the number of hits obtained for each pattern to build a feature for the classifier. While the underlying principle was the same, the indicative contexts as well as the input X to such patterns differed in the two tasks. In both cases we submitted the pattern instantiations to the web using the Google API.

For BioCreative, we built patterns for each entity X identified as a gene by an initial run of the tagger. If at least one of the patterns returned more than zero hits, the string was assigned a ‘web’ value for the Web feature. The classifier was then run again; this time incorporating the web feature. Using web-querying only on likely candidates for genes as identified by an initial run of the tagger was more efficient than using it on all words. However, this method does not contribute to improving recall.

In the BioNLP task, we experimented with a different approach. We built indicative contexts for each of the five classes to be recognised and for each word X which had a frequency lower than 10 as estimated from the British National Corpus (BNC) ¹ (Kilgarriff, 1997), we submitted the instantiation of each pattern to the Web. The pattern that returned the highest number of hits determined the feature value (e.g. “web-protein”, or “web-RNA”). If no hits were returned by any pattern, a value “O-web” was assigned. The same value was assigned to all words whose frequency was higher than 10.² This method proved less successful than the one used in our BioCreative system; it is unclear whether this is due to the method or to differences in the BioNLP task. In future work we will reproduce the same experiments on the two datasets in order to answer this question.

3.2.4. Gazetteer

Our gazetteer was compiled from lists of gene names from biomedical sites on the Web (such as Locus Link) as well as from the Gene Ontology and the data provided for BioCreative Tasks 1A and 1B. The gazetteer was cleaned by removing single character entries (“A”, “1”), entries containing only digits or symbols and digits (“37”, “3-1”), and entries containing only words that could be found in the English dictionary CELEX (“abnormal”, “brain tumour”). The final gazetteer contained 1,731,581 entries.

3.2.5. Frequency

We sought to incorporate information on frequency primarily as a way to weight information from external resources and to a lesser extent to indicate independently which tokens might be names. Because more frequent words are more likely to be ambiguous and less frequent words are far less likely to be ambiguous, we assumed that information from external resources would be of greater use for low frequency words. We therefore assigned to each

¹The BNC is a 100-million word corpus taken from a wide variety of sources.

²Using yet another value for words with higher frequency did not improve the tagger’s performance.

word a frequency category corresponding to the number of times the word was seen in a corpus. For BioCreative the corpus used was the BioCreative training data. For BioNLP, we improved on this by using counts from the BNC. We found that the frequencies obtained from the BNC were more intuitive than frequencies from a medical corpus.

3.3. Postprocessing

For BioCreative, we found that many of our errors stemmed from gene boundaries and addressed this issue in several ways. We removed genes containing mismatched parentheses from our results. We also found that we obtained different boundaries when we ran the classifier forwards versus backwards (reversing the order of the words) and obtained a significant improvement by simply combining the two sets of results and then keeping only the shorter entity in cases where one entity was a substring of another. We found that this postprocessing was highly valuable and added approximately 1% to our f-score. For BioNLP, we found that postprocessing was not useful because running the classifier forwards produced poor results and because mismatched parentheses were less of a problem.

4. Results and Analysis

	Precision	Recall	F-Score
gene/protein	82.8	83.5	83.2

Table 2: Results for BioCreative

	Precision	Recall	F-Score
protein	77.4	68.5	72.7
DNA	66.2	69.6	67.9
RNA	72.0	65.9	68.8
cell line	59.0	47.1	52.4
cell type	62.6	77.0	69.1
Overall	71.62	68.6	70.1

Table 3: Results for BioNLP

The performance of the system in both tasks is shown in Tables 2 and 3; the system gets an overall f-score of 83.2 for the BioCreative NER task and 70.1 for the BioNLP task. Our system compared well with other systems in the BioCreative evaluation; results from the BioNLP evaluation are forthcoming. Comparison to other results published on GENIA NE subsets is difficult because groups choose different subsets of GENIA entities and often evaluate on private corpora. (Shen et al., 2003) reports an f-score of 66 on a 24-NE task using version 3 of GENIA to evaluate. (Collier et al., 2000) and (Koichi and Collier, 2003) attempt a 10-NE task using a private corpus to evaluate and report f-scores of 74 and 73. We have analyzed our sources of error for both BioCreative and BioNLP in depth in (Dingare et al., 2004) and (Finkel et al., 2004); these include a large percentage of boundary errors (over 30% for both tasks), a smaller number of errors due to coordination, and some errors due to acronyms and tokens whose orthographic form might suggest they were entities but were in

fact measures or belonged to other entity categories; also a number of errors due to low-frequency words or words not encountered in the training data. However, we would like to focus here on the quality of training and evaluation data as a key factor leading to low performance.

The 13-point discrepancy between performance in BioCreative and BioNLP might be partially explained by the varying task difficulty: BioNLP requires recognition of 5 entities while BioCreative requires only 1; BioNLP also requires disambiguation of systematically ambiguous gene and protein names. However, task difficulty does not appear to be the primary factor leading to lower performance. To demonstrate this, we evaluated the system’s performance on the BioNLP data for the task of identifying a single class. When we eliminated the “cell line” and “cell type” categories and combined the “DNA” “RNA” and “protein” categories into a single class, we obtained an f-score of 74.4. This figure is substantially below the performance of 83.2 obtained for the roughly equivalent “NEWGENE” class in BioCreative. Rather than task difficulty, lower performance in BioNLP stems from higher inconsistency in the annotation of the BioNLP data. In saying this, we refer not only to errors in the evaluation data which resulted in lower scores but equally to inconsistencies in the training data which caused the system to learn incorrect patterns. Two of the authors independently reviewing 50 of the system’s errors found that 34-35 of these could be attributed to inconsistent annotation of the training or evaluation data. We are not biologists; we based our judgments of inconsistency on similarity of context. However, the example pairs we list below are so similar that we do not think the annotation inconsistencies are due to biological subtleties.

4.1. Data Annotation

Approximately one-third of the system’s errors were due to highly variable annotation of frequent terms such as *lymphocyte*, *T cell* and *B-cell*; these were variously annotated as “cell type” and as “O” (i.e. not in an entity). In example (1) below from the evaluation data our system labelled *lymphocytes* as a “cell type” and was penalized for a FP. However, our annotation is consistent with example (2) which appeared only two sentences later in the evaluation data; *lymphocytes* is annotated as a “cell type” there.

- (1) ...content of cAMP was also decreased in lymphocytes by 33% .
- (2) ...simultaneous alteration in the cAMP content was observed in *lymphocytes*.

Parallel problems occurred with the frequent terms *hormone* and *receptor* which were variously annotated as “protein” and “O”. In example (3) from the evaluation data our system labelled *receptors* as “O” rather than “protein” and was penalized for a FN; however our annotation mirrors example (4) which appeared in the training data.

- (3) Concentration of the *receptors* to 1.25 (OH) 2D3 was elevated up to 39.7 fmolemg after I week...
- (4) Concentration of receptors of hormonal form of 1 , 25 (OH) 2D3 was found to be minimal...

In a smaller proportion of cases entities were variably annotated either “DNA” or “protein”. In example (5) below which appeared in the evaluation data *kappa B enhancer* was labelled as “protein” while in example (6) which appeared in the training data it was labelled as “DNA”. Variation in labelling between “DNA” and “protein” also occurred with *enhancer elements*.

- (5) These kappa B-specific proteins...interact with the functional *kappa B enhancer* present in the IL-2R alpha promoter .
- (6) ...nuclear NF-kappa B is necessary to activate the *kappa B enhancer*...

Inconsistent annotation of premodifiers also caused a small number of errors. In examples (7), (9), and (11) which appeared in the evaluation data, the modifiers *human*, *inducible*, and *unrearranged* were included in the entities “DNA”, “protein”, and “DNA”, respectively, while in the parallel examples (8), (10), and (12) which appeared in the training data, they were excluded. Our system left out the modifiers as in the training data and was penalized for both FPs and FNs.

- (7) Kappa B-specific DNA binding proteins: role in the regulation of *human interleukin-2 gene* expression.
- (8) Instead , signal transduction to the human *IL-2 gene* became disrupted .
- (9) Mutation of a kappa B core sequence...blocks the specific binding of two *inducible cellular factors*.
- (10) [Sequence analysis revealed] several putative binding sequences for inducible *transcription factors*...
- (11) Different fragments of *unrearranged human variable region*...were used for...in vitro transcription....
- (12) ...hGATA-3 may be involved in the regulation of the unrearranged *TcR delta gene* expression....

Some cases of inconsistent annotation were due to cancer terms such as *neoplasm*, *tumor*, and *carcinoma* which were annotated either as “cell type” or “O”; we assume that this is because these terms are ambiguous between cell types and disease names.

- (13) ...the authors studied specimens of *breast carcinomas* from 60 consecutive female patients.
- (14) Inflammatory infiltrates were analysed in tissue sections of 76 breast carcinomas...

There was also uncertainty as to whether gene systems, core sequences, and stretches of DNA described by numerical location (e.g. -206 to -195) should count as “DNA” entities. Finally, there was highly variable annotation of coordination.

Overall, the quality of data in the BioCreative evaluation appeared to be significantly higher and did not feature the systematic inconsistencies of the BioNLP data (keeping in mind that the BioCreative annotation task was also significantly easier). BioCreative’s innovation of enumerating several alternate correct boundaries reduced spurious

boundary errors. However, there were some inconsistencies in the BioCreative data as well. In a few cases organism names appearing in prepositional phrases after gene names were annotated as if they were premodifiers (as in (15)) while in other cases they were not (as in (16)).

- (15) Transcriptional regulation of *SUP35* and *SUP45* in *Saccharomyces cerevisiae*
- (16) Expression of the...protein Bax under the control of a *GAL10 promoter* in *Saccharomyces cerevisiae* resulted in...

The annotation of *mutations* was also inconsistent - the participants were given instructions not to annotate *mutations* as genes and were given the example *p53 mutations*; but in the training data there were 25 instances of *mutations* annotated as genes, including *p53 mutations*.

4.2. Improving Biomedical Annotation

That the task of biomedical NER is more difficult than NER in the traditional newswire domain (with its standard entities of “PERSON”, “LOCATION” and “ORGANIZATION”) is obvious from the numbers; the highest score in the CoNLL 2003 NER task (Sang and De Meulder, 2003) (which used the same scoring metric as BioNLP and BioCreative) was 88.8%, five points higher than the highest score in BioCreative, and 18 points higher than our score in BioNLP. What must be noted is that the difficulty of the domain has an effect both on the annotation of the data as well as on the performance of the system. In a difficult domain where language is convoluted and names are long and complex, data annotation is more difficult. This is demonstrated by results on interannotator agreement – while interannotator agreement for the MUC-7 NER task in the newswire domain was measured at 97% (Marsh and Perzanowski, 1998), the few studies of interannotator agreement in the biomedical domain have shown interannotator agreement to be substantially lower, with f-scores in the range of 0.87 (Hirschman, 2003) to 0.89 (Demetriou and Gaizauskas, 2003). In order to accurately represent the state of progress in biomedical NER, evaluations must focus as much on improving biomedical data annotation as on improving systems. We note that while the use of annotation guidelines has become standard practice particularly for complex annotation tasks, the annotation of the BioCreative data did not use annotation guidelines. We also know of no guidelines used in the annotation of the GENIA data used in the BioNLP task. The adoption of annotation guidelines in a domain notorious for its complexity and where interannotator agreement is known to be low seems to be a promising direction for improvement.

Annotation guidelines must address the proper annotation of premodifiers, constructing rules to distinguish the premodifiers that are necessary to annotate. They must also specify how to annotate coordinated entities, distinguishing between the varieties of coordinations. Next, they must establish whether to annotate high-level categories. It may be that the variability in the annotation of words like *receptor* and *hormone* was due to the fact that receptors and hormones are types of protein containing thousands of

instances. Finally, annotation guidelines must decide ambiguous cases of class membership such as whether DNA sequences are examples of “DNA” entities and whether tumors are “cell types”.

5. Conclusions

We have presented a machine learning system for biomedical NER and presented its performance in the two biomedical NER evaluations to date. Our system’s rich feature set including exhaustive use of local features and a variety of external resources leads to state-of-the-art performance. Our system also adapts rapidly to new NE sets as illustrated by our adaptation to the BioNLP task.

Unfortunately, state-of-the-art-performance in biomedical NER continues to lag behind the high-eighties figures that the field has come to expect. The BioNLP organizers may have had this gap in mind when they emphasized that participants should focus on deep knowledge sources such as coreference resolution and use of dependency relations over “widely used lexical-level features (POS, lemma, orthographic, etc.)”. However, both BioNLP and BioCreative showed that external resources led to improvements of only 1-2%. Our error analysis showed that consistent annotation might have led to a 70% reduction in error rate. While the proper exploitation of external resources and deep processing remains an avenue to be explored, we believe it cannot compare to the gains that might result from consistently annotated data. The challenge for future evaluations is to use and publish annotation guidelines, to measure and report figures for interannotator agreement, and to pursue improvements in annotation of biomedical data alongside improvements in systems.

6. Acknowledgments

This work was supported by a Scottish Enterprise Edinburgh-Stanford Link Grant (R36759) as part of the SEER project and by the National Science Foundation under the Knowledge Discovery and Dissemination program.

7. References

- Christian Blaschke, L. Hirschman, and A. Yeh, editors. 2004. *Proc. of the BioCreative Workshop*, Granada, March. http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/handout/.
- Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *ANLP 6*, pages 224–231.
- Nigel Collier, C. Nobata, and J. Tsujii. 2000. Extracting the names of genes and gene products with a hidden markov model. In *Proc. of CoLing*, pages 201–207.
- Nigel Collier, J. Kim, Y. Tateisi, T. Ohta, and Y. Tsuruoka, editors. 2004. *Proc. of the International Joint Workshop on NLP in Biomedicine and its Applications*, Geneva, August. to appear.
- James R. Curran and S. Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proc. of the Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 164–167, Edmonton, Canada.
- George Demetriou and R. Gaizauskas. 2003. Corpus resources for development and evaluation of a biological text mining system. In *Proc. of the Third Meeting of the Special Interest Group on Text Mining*, Brisbane, Australia, July.
- S. Dingare, J. Finkel, M. Nissim, C. Manning, and B. Alex. 2004. Exploring the boundaries: Gene and protein identification in biomedical text. In *Proc. of the BioCreative Workshop*.
- Jenny Finkel, S. Dingare, H. Nguyen, M. Nissim, and C. Manning. 2004. From syntax to the web. In *Proc. of the Intl. Joint Workshop on NLP in Biomedicine and its Applications at CoLing 2004*, Geneva, Switzerland, August.
- K. Fukuda. 1998. Toward information extraction: Identifying protein names from biological papers. In *Proc. of the Pacific Symposium on Biocomputing*, pages 705–716.
- Gregory Grefenstette. 1999. The WWW as a resource for example-based MT tasks. In *Proc. of ASLIB’99 Translating and the Computer 21*, London.
- Lynette Hirschman. 2003. Using biological resources to bootstrap text mining. Presentation to the Massachusetts Biotechnology Council Informatics Committee.
- Jun’ichi Kazama, T. Makino, Y. Ohta, and J. Tsujii. 2002. Biomedical name recognition: Tuning support vector machines for biomedical named entity recognition. In *Proc. of the ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*, pages 1–8.
- Frank Keller and M. Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.
- A. Kilgarriff. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155.
- Dan Klein and C. Manning. 2003. Accurate unlexicalized parsing. *ACL 41*, pages 423–430.
- Dan Klein, J. Smarr, H. Nguyen, and C. D. Manning. 2003. Named entity recognition with character-level models. In *CoNLL 7*, pages 180–183.
- T. Koichi and N. Collier. 2003. Bio-medical entity extraction using support vector machines. In *Proc. of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*.
- K. Markert, M. Nissim, and N. Modjeska. 2003. Using the web for nominal anaphora resolution. In R. Dale, K. van Deemter, and R. Mitkov, editors, *Proc. of the EACL Workshop on the Computational Treatment of Anaphora*, pages 39–46.
- E. Marsh and D. Perzanowski. 1998. MUC-7 evaluation of IE technology: Overview of results. In *Message Understanding Conf. Proc., 7-proceedings/marsh_slides.pdf*.
- Andrew McCallum, D. Freitag, and F. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proc. of the 17th International Conf. on Machine Learning*.
- Andrei Mikheev, M. Moens, and C. Grover. 1999. Named entity recognition without gazetteers. In *Proc. of EACL99*, pages 1–8, June.
- C. Nobata, N. Collier, and J. Tsujii. 1999. Automatic term identification and classification in biology texts. In *Proc. of the 5th NLPRS*, pages 369–374.
- Tomoko Ohta, Y. Tateisi, H. Mima, and J. Tsujii. 2002. GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of HLT 2002*.
- Erik F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proc. of CoNLL-2003*, pages 142–147.
- Ariel Schwartz and M. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, Kauai, Jan.
- Dan Shen, J. Zhang, G. Zhou Jian Su, and C. Tan. 2003. Effective adaptation of hidden markov model-based named entity recognizer for biomedical domain. In *Proc. of the ACL-2003 Workshop on Natural Language Processing in Biomedicine*.

Protein Name Tagging Guidelines: Lessons Learned

Inderjeet Mani, Zhangzhi Hu, Seok Bae Jang, Ken Samuel[†],
Matthew Krause, Jon Phillips and Cathy Wu

37th and O Sts, NW
Georgetown University
Washington, DC 20057

[†]The MITRE Corporation
7515 Colshire Drive, McLean, VA 22102

Abstract

This paper is motivated by the need to establish common definitions of the problem of protein name tagging. We describe the lessons learned in developing a set of guidelines and present the first set of inter-coder results, viewed as an upper bound on system performance. The guidelines, annotated datasets, along with automatic tools, are available for research use.

1 Introduction

With the enormous quantity and variety of high-throughput data being generated in the post-genome era, one of the major challenges in managing biological knowledge is to provide timely, accurate, and consistent annotation of the biological databases such as primary DNA (GenBank) and protein sequence databases (UniProt) and many other secondary databases. Of particular value is annotation derived from experimentally verified data published in the scientific literature. However, the amount of such literature-based and manually-curated annotation is rather limited due to the laborious nature of knowledge extraction from the literature. Interest in information extraction from the biomedical literature is motivated by the need to speed up the creation of structured databases representing the latest scientific knowledge about specific objects such as proteins and genes. This has resulted in

natural language processing technologies being utilized for biological literature mining and information extraction (Hirschman *et al.*, 2002).

We discuss here our experience in developing resources for one particular problem area¹, that of extracting protein names from MEDLINE abstracts. This task is fundamental to several other biological literature mining tasks, including the development of protein name ontologies and extraction of protein annotations (such as function and protein-protein interaction) from literature.

2 The Problem

Protein names show considerable variation because of the existence of multiple naming conventions. Researchers may name a newly discovered protein based on its function, sequence features, gene name, cellular location, molecular weight, or other properties, as well as abbreviations and acronyms. For example, the EphB2 receptor, a protein involved in signaling in the brain, was initially referred to as ‘Cek5’, ‘Nuk’, ‘Erk’, ‘Qek5’, ‘Tyro6’, ‘Sek3’, ‘Hek5’, and ‘Drt’ before being standardized as ‘EphB2’ (Nature 1999). Potential standardization based on publishing guidelines and community consensus on naming are hard to uniformly enforce. Moreover, there are proteins whose status is tentative, and there is of course also a vast amount of legacy data.

Unfortunately, the previous research in protein and gene name tagging has been hampered in several ways. Some systems

¹ This research was supported by the National Science Foundation (ITR-0205470).

distinguish between protein and gene names, others don't, but the criteria for specifying when a protein or gene name should be tagged are not discussed. Thus, in addition to the lack of common datasets, it becomes very difficult to compare systems if one is unsure if they are addressing the same problem. By using common *tagging guidelines*, it becomes possible for groups to share tagged data, compare automatic tagging results, and in general advance the field of biological information extraction. Also, *inter-coder reliability* is hardly ever reported (a notable exception is (Hatzivassilioglou et al. 2001)). As a result, one has no real sense of how difficult the task is, and as to how well the machine is faring relative to the upper bound of human performance.

The BioCreAtIvE evaluation is motivated by similar concerns, and is a very positive step that should address some of these issues. We believe strongly that our resources and approach can be leveraged in such evaluations.

3 Tagging Guidelines v1

3.1 Focus

Our first set of guidelines was relatively ambitious. We began with the assumption that it was crucial to annotate references to protein objects (including protein complexes and sets of protein objects), rather than simply annotating the protein names. References to genes, gene promoters, mutant genotypes, etc., were therefore not tagged. Thus, "HypA" was tagged as a protein, while "hypA", which refers to a gene, was not. **Ambiguity** between genes, proteins and genotypic strains was addressed by specific conventions.

3.2 Tag Type

We defined three tag types: (i) *<protein>* as a generic tag for most protein objects, including protein complexes (e.g. "pyruvate dehydrogenase complex"); (ii) *<acronym>* to tag acronyms or abbrevia-

tions; (iii) *<array-protein>* to tag a list of proteins as a whole (e.g., "FGF-1, -2, -4, -5, and -7").

3.3 Tag Extent

Our rules for tag extent were reasonably complex. A name was assumed to be made up of a pre-modifier chunk, a head, and a post-modifier chunk. Protein names were not tagged when used as modifiers for non-protein entities (e.g., "elastase I promoter"). When the post-modifier in a name expressed a part-of relation (subunits or chains of a complex), the name was tagged as a whole, e.g., *<protein> subunit of NADH dehydrogenase (complex I)</protein>*. However, if the part referred to a subregion of a protein or a polypeptide such as "c-terminal tail of the hLHR", only the head "hLHR" was tagged. Other rules were defined for kind-of and member-of relations, as well as various other cases.

4 Data and Annotation Procedure

We created two sets of 300 abstracts (called ABS1 and ABS2), each corresponding to 300 PIR-NREF protein entries that were randomly picked from about 5000 entries with curated information from high quality underlying databases such as PSD (Protein Sequence Database), SGD (*Saccharomyces* Genome Database) and Locus-Link.

ABS1 was tagged by hand by one coder using MITRE's Alembic Workbench (Aberdeen et al. 1995). The human coder tagged nearly 3300 protein names in them. This experience provided a basis for developing a formal set of guidelines. ABS2 was then tagged according to the guidelines by three human coders using the Workbench. A1 was a paper author, while the others were biologists otherwise unconnected with this project.

5 Assessment of v1

Coders	Correct	Precision	Recall	F-measure
A1-A2	3091	0.750	0.748	0.749
A1-A3	2766	0.8250	0.669	0.739
A3-A2	2474	0.6	0.738	0.662
Average		0.67	0.771	0.716

Table 1: Inter-Coder Reliability on protein tags v1

The inter-coder reliability metrics computed by a MUC-class named entity scorer used in the DARPA TIDES program is shown in Table 1. The scorer is strict in that a candidate name and a reference name match (such a match is labeled “Correct” in our tables) if and only if their respective text extents have exactly the same characters at exactly the same positions in the text.

Kappa is often used to measure inter-coder reliability on classification tasks, but its extension to named entity extent is less clear. We consider each word position in the abstract, and compare whether the word at that position is a component of a protein name or not across coders. In addition to ignoring the boundary between contiguous protein names, this measure is generous and could give artificially high scores because most words are not components of protein names, though chance agreement can be high. A related measure is used in (Marcu et al. 1999) for computing Kappa on discourse spans. This method gives a Kappa of 0.80.

The ambitious **focus** on protein objects was a major reason for disagreement. Many of the cases of disagreement involved **ambiguity** of names that could refer to either genes or proteins. Moreover, many context-specific protein objects were tagged by some coders even when they were very generic

(e.g., protein, enzyme) or meaningless when taken out of context (e.g. “E1 alpha”).

We next consider **extent**. While the maximum protein name length was 12 words, about 93% of the tags were three words or less, and 86% of the tags were two words or less, and agreement on these was much higher. Coders were inconsistent in annotating pre- and post-modifiers and morphological affixes at the boundary of a name, and also in incorporating trailing punctuation in the tag. Nearly half such tags were off by just one word.

Finally, consider **tag types**. Acronym tagging achieved a 0.85 F-measure, but here the guidelines were not consistently followed. The array-protein tags were very hard to annotate (0.15 F-measure). This was because they were not clearly defined in the guidelines, e.g., a list of protein objects may or may not share a common core term.

Finally, coders showed **fatigue**, and often missed tagging multiple occurrences of the same protein name.

6 Tagging Guidelines v2

The above sorts of considerations led us to revise the guidelines, as discussed next.

6.1 Focus

In the previous guidelines, when a protein name was followed by a non-protein object (e.g., “elastase I gene promoter”), the protein name (e.g., “elastase I”) was not tagged. This was because of the focus on protein objects. In the modified guidelines, we defined the tagging targets as protein named entities (full names, acronyms or other symbolic names) used in the literature to describe proteins, or protein-associated or -related objects, such as domains, pathways, expression, or gene. Thus, in the new guidelines, we have *<protein>elastase I</protein> gene promoter*, etc.

6.2 Tag Types and Extent

In the revised guidelines, we used only two types of tags: *<protein>* and *<long-form>*. The *<long-form>* tag is de-

signed to optionally extend the boundaries of <protein> tag when the name boundary is difficult to determine, thereby improving inter-annotator consistency. The long-form is only used in two situations (more details are at our website):

1. Organism names preceding a protein name may or may not be part of the protein name. For example, the species name is tagged as part of the protein name if the protein name contains an acronym abbreviating the species name, e.g., <protein>human growth hormone (hGH)</protein>, but <long-form>human <protein>IGF-II</protein></long-form>.
2. When several protein entities share common terms, there may be only one name entity that can be easily tagged. We tag such an entity as a protein, while the list of entities together are tagged as a long-form, e.g., <long-form><protein>CSN subunits 4</protein>, 5, 6</long-form>.

7 Assessment of v2

Coders	Correct	Precision	Recall	F-measure
<protein>				
A1-A3	4497	0.874	0.852	0.863
A1-A4	4769	0.884	0.904	0.894
A3-A4	4476	0.830	0.870	0.849
Average		0.862	0.875	0.868
<long-form>				
A1-A3	172	0.720	0.599	0.654
A1-A4	241	0.837	0.840	0.838
A3-A4	175	0.608	0.732	0.664
Average		0.721	0.723	0.718

Table 2: Inter-Coder Reliability F-measure (v2)

The results on inter-coder reliability using the revised guidelines are much better. We present results for F-measure in Table 2 with three coders² on ABS2. The corresponding Kappa scores are shown in Table 3.

Coders	Kappa
<protein>	
A1-A3	0.899
A1-A4	0.930
A3-A4	0.892
3-way	0.932
<long-form>	
A1-A3	0.657
A1-A4	0.819
A3-A4	0.662
3-way	0.766

Table 3: Inter-Coder Reliability Kappa (v2)

8 Related Work

Other work on inter-coder reliability comes from (Hatzivassiloglou et al. 2001), who had 3 annotators manually classify 550 terms found in 15 full-text articles from PubMed as gene, protein, mRNA, ambiguous, or wrongly extracted. They found 77.58% pairwise agreement and 69.27% three-way agreement.

We now compare our annotated corpus with the GENIA corpus version 3.0.2 (Ohta et al. 2002). The latter is a 2000-abstract corpus of biological literature compiled from the MEDLINE database and tagged with a set of hierarchical semantic classes. The GENIA corpus is focused on biological reactions concerning transcription factors in human blood cells, with the MeSH terms “human”, “blood cell” and “transcrip-

² Note that the coders A1 and A3 were involved in v1 as well.

tion factor” used as criteria for selecting abstracts.

The corpus has clearly a different focus from ours. Our corpora were chosen from the curated NREF database entries, which are not biased towards any particular area of biology, thus providing greater diversity in protein names for a given sample size. In addition, of course, our focus is on tagging protein names, a fundamental problem in automatically extracting experimental information of proteins from literature to assist protein database annotations.

The GENIA ontology classes corresponding to our protein name entities are “protein complex”, “individual protein molecule”, “subunit of protein molecule”, and “peptide” (here, we exclude peptides as only naturally occurring peptides map to protein name objects, not artificial synthetic peptides). Based on our mapping, both corpora have a similar percentage (about 22%) of distinct protein names.

9 Resources

A dictionary of 691,000 protein names was compiled from PIR NREF entries. A case-insensitive exact matching of longest matching entries achieved an F-measure of 0.412 (0.372 Precision, 0.462 Recall) on ABS2. When used for preprocessing before coding, we found the dictionary lookup helpful with **standardization** and **extent**. It should also help with the **fatigue** problem, and thus could considerably further improve inter-coder reliability.

We have also developed several automatic taggers (also available if desired) based on machine learning, which currently perform at about .59 F-measure, tested on both ABS2 as well as the 2000-abstract GENIA corpus version 3.0.2 (Ohta et al. 2002), with the latter being based on a mapping of GENIA tags to ours.

These results compare with a .40 F-measure for KEX (Fukuda et al. 1998) on ABS2, and are comparable with other work on GENIA. In the (Hatzivassiloglou et al.

2001) study, their best automatic taggers were at 7-14% below human performance.

Our guidelines and annotated data (600 abstracts in all) are available to the community, along with a general corpus study and more detailed results (see compilingone.georgetown.edu/~prot/). Our resources will be discussed along with more detailed results at the symposium.

References

- Aberdeen, J. Burger, J. Day, D. Hirschman, L., Robinson, P., and Vilain, M. 1995. *MITRE: Description of the Alembic System as used for MUC-6*. In Proceedings of the Sixth Message Understanding Conference (MUC-6), 141-155.
- Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. 1998. *Information extraction: identifying protein names from biological papers*. Proceedings of PSB'98.
- Hatzivassiloglou, V., Duboue, P. A., and Rzhetsky, A. 2001. *Disambiguating proteins, genes, and RNA in text: A machine learning approach*. Bioinformatics, 17, Supplement 1, S97-S106.
- Hirschman, L., Park, J.C., Tsuji, J., Wong, L., and Wu, C.H.. 2002. *Accomplishments and challenges in literature data mining for biology*. Bioinformatics Review, 18, 12, 1553-1561.
- Marcu, D., Romera, M., and Amorrortu, E. 1999. “Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues”. *Proceedings of the Workshop on Levels of Representation in Discourse*, 71-78.
- Nature (editorial opinion).1999. *Wanted: A new order in protein nomenclature*. Nature, 401, 411, 30 September 1999.
- Ohta, T., Tateisi, Y., Kim, J-D and Tsuji, J. 2002. *The GENIA Corpus: an Annotated Corpus in Molecular Biology Domain*. In the Proceedings of the 10th International Conference on Intelligent Systems for Molecular Biology (ISMB 2002) poster session.

Ontology-Based Interactive Information Extraction from Scientific Abstracts

Milward, David¹, Bjärelund, Marcus², Hayes, William³, Maxwell, Michelle⁴, Öberg, Lisa², Tilford, Nick⁴, Thomas, James¹, Hale, Roger¹, Knight, Sylvia¹ and Barnes, Julie⁴

¹ Linguamatics Ltd., St. John's Innovation Centre, Cambridge, CB4 0WS, UK

² AstraZeneca R&D Mölndal, SE-431 83 Mölndal, Sweden

³ AstraZeneca R&D Boston, 35 Gatehouse Drive, Waltham, MA 02451, USA

⁴ BioWisdom Ltd, Babraham Hall, Babraham, Cambridge, CB2 4AT, UK

david.milward@linguamatics.com, marcus.bjareland@astrazeneca.com,
william.s.hayes@astrazeneca.com, julie.barnes@biowisdom.com.

Abstract

Over recent years, there has been a growing interest in extracting information automatically or semi-automatically from the scientific literature. This paper describes a novel Ontology-based Interactive Information Extraction (OBIIE) framework and a specific OBIIE system. We describe how this system enables life scientists to make *ad hoc* queries similar to using a standard search engine, but where the results are obtained in a database format similar to a pre-programmed information extraction engine. We present a case study where the system was evaluated for extracting cofactors from EMBASE and MEDLINE.

Introduction

Information Retrieval (IR) systems are designed to find the highest ranked documents that match a user query, such as a set of keywords. This contrasts with Information Extraction (IE) which returns relationships, e.g. a table of protein-protein interactions or gene-disease relationships, rather than a ranked set of documents. Traditionally these two techniques have been seen as very different: IR is all about finding documents, IE about finding facts within documents. However, from a user perspective the difference is not so great. The users' aim is primarily to get to information as fast as possible: if a system can get them directly to relevant sentences then this will save time. If the system can also get them to structured results, appropriate sorted, this can save further time. Interactive Information Extraction is a new concept which combines the interactive querying style of a web search engine with the structured output that is provided by standard IE. This allows scientists to refine queries and explore a set of texts in a similar way to web search, but with the possibility of much more precise search and results.

In the Linguamatics I2E System, the user can start with a standard search for words within a document, then refine this to require the words to be in the same sentence, or in a particular linguistic pattern. The results are output as HTML tables or in a format suitable for database entry. The system pre-indexes documents such as MEDLINE abstracts to allow fast querying. All linguistic processing is done prior to indexing, including "tokenisation" to split a string of characters into individual words, "sentence splitting" to recognise sentence ends, "tagging" to recognise parts of speech such as nouns or verbs, and "chunking" to group words into meaningful units according to their parts of speech.

The OBIIE Framework and System

Conventional IE systems typically allow extraction of syntactic classes (nouns or verbs) and a few unstructured semantic classes, so called named-entities such as proteins, diseases, or amounts.¹ Until recently there has been little use of richer domain knowledge within such systems. In OBIIE, ontologies provide that domain knowledge, enabling the users to interact with the system on a conceptual level without having to know all possible synonyms for a concept. The relationships “part-of” or “is-a” provide a basic taxonomy allowing the user to choose particular concepts or families of concepts (see Figure 1). The user can now construct queries using the many thousands of different classes found in typical ontologies.

A scientist may start with a query such as two keywords in the same document (equivalent to standard keyword search), then refine this to look for two *classes* e.g. a disease and a protein (co-occurrence within a document), then refine this further to require co-occurrence within a sentence, and finally refine this to look for the protein and disease in a particular syntactic configuration. Incorporation of linguistic constraints is up to the user, and is justified by an increase in precision that is enough to balance any decrease in recall. For example, a query for the word “RAF” followed by the word “phosphorylate” would be improved by putting “phosphorylate” within a so-called “verb group”. This would then match the text string “Raf has been shown to phosphorylate”, without returning the larger number of false hits you would have got by allowing Raf within e.g. 5 words of phosphorylate.

The first OBIIE system was constructed by putting together ontologies from BioWisdom with the I2E System. The ontologies incorporated within OBIIE express ‘is-a’, ‘is-a -part-of’ (i.e. taxonomic) relationships as well as the ‘has synonym’ relationship. Ontological concepts can be selected from up to 25 concept types, ranging from genes, proteins, tissues, cells, clinical disorders, symptoms, processes, pathways, drugs, adverse effects, and techniques, technologies etc. Where appropriate, ontologies can be used as species-specific structures, or as master species-independent ontologies.

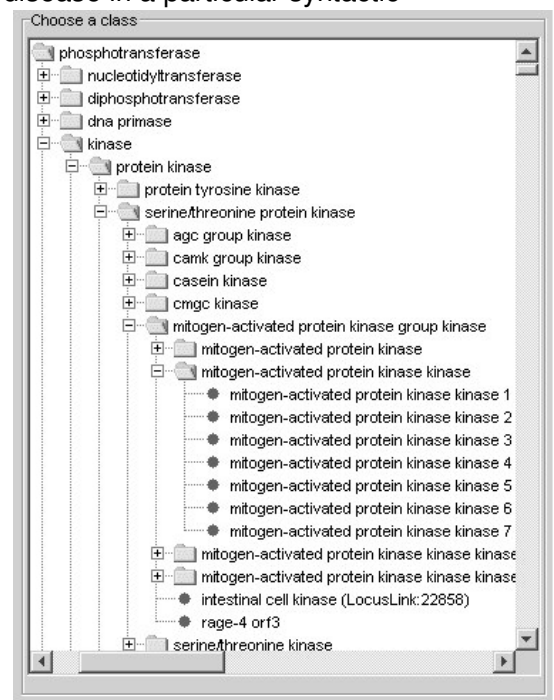


Figure 1: Protein Ontology

As well as defining patterns on the fly, it is possible to reuse existing patterns or pattern templates. Patterns can be organised hierarchically to provide relationship ontologies, so that a user interested in interactions between drugs and proteins can choose the family of “drug interaction relationships”, and obtain alternative phrasings such as the verbs “activates”, “inhibits”, “blocks” (and their morphological variants) or phrases such as “is agonist for”.

¹ Named entities may be closed classes formed from an enumerated list e.g. a list of names, or open classes recognised via patterns (e.g. protein spotting routines such as Fukuda et al. 1998).

Case Study : Text-Mining for Nuclear Receptor Cofactors

In this evaluation we compared the use of two methods. The first consisted of finding a set of abstracts, reading the abstracts, and extracting cofactors from these. The second used the OBIIE tool over a larger set of abstracts, extracting the results, and then filtering these by hand. We were particularly interested in the relative recall between the two methods, and the speed of each method.

Nuclear receptors (NRs) are ligand-dependent transcription factors that typically recruit protein complexes (cofactors) to enhance or repress transcription of target genes. It is believed that several phenomena such as level of transactivation and tissue specificity of NRs heavily depend on the specific recruited cofactors. Since NRs are very important drug targets (18 of the 48 known human NRs are targets for registered drugs) the amount of literature on these proteins is rapidly increasing. The aim of this case study was to generate a comprehensive and annotated lists of cofactors for three NRs: Androgen Receptor (AR) and Liver X Receptors (LXR) α and β . AR abstracts were used for "training" (i.e. query tuning and ontology refinement) and LXR abstracts for testing. The project validated the use of ontology- and linguistic-based text-mining against the previous best practice for obtaining sets of cofactors, based on published or manually generated lists.

For AR we constructed a secondary corpus of 7748 abstracts from MEDLINE and EMBASE by fine-tuning synonym choices. We used two sources for the manual list of cofactors for AR: A list of AR-interacting proteins compiled by Dr. Lenore Beitel (Beitel, 2002), and a list constructed by manually examining a subset of abstracts from the secondary corpus containing ~300 abstracts. The abstracts underlying this sub-corpus were further analyzed in several iterations with regard to sentences containing information on cofactor recruitment. When we reached sufficient recall (~90%) of extracted relationships we stopped the iterations. The relationship ontology includes various alternative phrasings and words within larger patterns. Compiling out the embedded alternatives gives a relationship ontology representing 188 distinct patterns. The results of applying the OBIIE system with the relationship ontology are shown in Table 1. The two top reasons why we failed to reach 100% recall was that the secondary corpus did not contain any abstracts with a reference to those symbols (in fact, no abstracts in MEDLINE or EMBASE contained co-occurrences of those symbols and AR), and that references spanned several sentences. Cross-sentence patterns were not used as they severely affected precision.

Table 1. Results for extraction of NR cofactors.

NR	# Abstracts in secondary corpus	# Abstracts retrieved by OBIIE	# Cofactors manually retrieved	# Cofactors retrieved by OBIIE	#Cofactors in found in total
AR	7748	564	101	100	110
LXR α/β	N/A	68	9	9	10

Instead of cross-validating the results on the AR secondary corpus, we applied the same query, with the AR synonyms replaced by a selected number of LXR α and β synonyms to the whole of MEDLINE (i.e. we did not construct a secondary corpus in this case). The manual list was generated by manual examination of 240 LXR abstracts, but the relationship ontology was not changed based on any of those abstracts. The results of applying OBIIE to extract LXRs from the whole of Medline (the set of Medline abstracts as of Nov 25th 2003) are displayed in Table 1. One cofactor was missed by the OBIIE system due to a missing pattern in the relationship ontology. The OBIIE system managed to retrieve a cofactor that manual curation missed. For LXRs, recall was 90%. Recall was calculated by dividing the number of cofactors found by

OBIIE, divided by the best figure available for the total number of cofactors (the total number of cofactors discovered by either method). Lists of the retrieved cofactors can be found at <http://www.linguamatics.com/obiie/>.

The results obtained by the OBIIE system are clearly satisfactory, not only for the recall achieved, but also for the amount of time saved by the automated process. Without risking too many human errors we estimate that one person can read 100 abstracts in a day. With the tabular output format from the OBIIE system, a domain expert can increase this by an order of magnitude, since only sentences with the NR symbol, a relationship phrase, and a cofactor symbol are displayed.

Related Work and Discussion

In this extended abstract, we have described the new method of Ontology-Based Interactive Information Extraction. There is prior work on the integration of ontologies with standard Information Extraction. For example, Mädche et al. (1999) discuss the use of ontologies as a way to provide information in a canonicalised format to a user or for input into a database, and Todirascu et al. (2002) use ontologies for identifying concepts. Although there are precedents for the use of ontologies in information extraction, there are a number of necessary features of any ontology applied in this way. The BioWisdom ontologies used in this case study incorporate a comprehensive list of synonyms, and have a detailed hierarchical structure allowing fine-grained concept distinctions. The domain specific nature of the ontologies also ensures that the synonyms are appropriate to the pharmaceutical domain.

Interactive Information Extraction has few precedents. There is some similarity with work on *question answering systems*. Here NLP and IR techniques are used to obtain the best answer to a question expressed in natural language. In contrast, I2E provides all results for a structured query. There is also some similarity with work which provides an interactive front end on top of the output of a fixed IE system (e.g. Gazaiskus et al. 2001). However, in these systems the result of a new query will always be a subset of the results provided by the original fixed IE query patterns. There are other systems which can be said to be positioned somewhere between document search (IR) and relationship search (IE), but they are typically designed for one specific task, e.g. looking for symbol co-occurrences in sentences. In contrast, in *Interactive Information Extraction* there is a natural gradation from document search, via search within sentences, to search for specific relationships within sentences. It is also possible to perform combined searches e.g. search for relationships, but only in documents containing a particular concept.

An examination of cofactors similar to the case study was performed by Albert et al. 2003 using a co-occurrence based approach. They searched for tri-occurrences, i.e. entity, relation, and entity within the same sentence. Recognition of entities was performed using finite state string matching automata (regular expressions) rather than using linguistic processing. They retrieved fewer cofactors, but it is not appropriate to compare the results directly, since they were working with a smaller corpus, theirs being extracted 10th Sep 2001 and ours during the fall of 2003 (Oct 21st for AR and Nov 25th for LXR).

The case study presented here highlights the power of OBIIE in performing systematic textual analysis. The use of synonyms, coupled with the interactive nature of the tool makes it very quick to engineer queries to accommodate the variety of linguistic forms that phrases take, and hence allows for high recall coupled with good precision. The results for cofactors were remarkably good, with new cofactors discovered by the OBIIE system that had not been in the original manually retrieved set. By exploiting the redundancy inherent in a large corpus, we were

able to use relatively specific patterns giving high precision, while still getting recall that rivals that achieved manually. Although the case study in this paper focused on extracting cofactors, the incorporation of other large biomedical ontologies covering areas of disorders, symptoms, tissues, cells, compounds, biological processes etc. makes this a flexible tool for use across all parts of pharmaceutical R&D.

Bibliography

- Albert, S., Gaudan, S., Knigge, H., Raetsch, A., Delgado, A., Huhse, B., Kirsch, H., Albers, M., Rebholz-Schumann, D. & Koegl, M. (2003) Computer-assisted generation of a protein-interaction database for nuclear receptors, *Molecular Endocrinology*, 2002-0424
- Beitel, L. (2002) List of AR-interacting proteins, available from *The Androgen Receptor Gene Mutations Database World Wide Web Server*: <http://ww2.mcgill.ca/androgendb/>
- Biowisdom, <http://www.biowisdom.com/>
- Blaschke, C., Hirschman, L., and Valencia, A. (2002) Information Extraction in molecular biology. *Briefings in Bioinformatics*. 3(2):1-12.
- Craven, M. and Kumlien, J. (1999) Constructing biological knowledge bases by extracting information from text sources. *ISMB*, AAAI Press, pp. 77-86.
- De Bruin, B. & Martin, J. (2002) Literature mining in molecular biology. *EFMI Workshop on Natural Language Processing in Biomedical Applications*, Baud, R. & Ruch, P. (eds.), Nicosia, Cyprus, 1-5. *EMBASE*, <http://www.embase.com>
- Fukuda, K, Tsunoda, T., Tamura, A., and Takagi, T. (1998) Toward information extraction: identifying protein names from biological papers. *Proc. PSB, Hawaii*, 3, 705-716.
- Gaizauskas, R., Herring, P., Oakes, M., Beaulieu, M., Willett, P., Fowkes, F. and Jonsson, A. (2001). *Intelligent Access to Text: Integrating Information Extraction Technology into Text Browsers*. *Proceedings of the Human Language Technology Conference (HLT2001)*, 189-193
- Hopkins, A.L. and Groom, C.R (2002) *The Druggable Genome*. *Nature Reviews Drug Discovery*, Vol. 1, 727-730.
- Humphreys, K., Demetriou, G. and Gezauskas, R. (2000) Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structure. *Pac. Symp. Biocomp.*, pp. 502-513.
- Jenssen, T.K, Laegrid, A., Komorowski, J., Hovig, Eivind, (2001), 'A literature network of gene for high-throughput analysis of gene expression', *Nature Genetics*, 28:21-28.
- Linguamatics Ltd., (2003) *Interactive Information Extraction: White Paper*, http://www.linguamatics.com/resources/white_paper_ie.html
- Maedche, A., Staab, S. and Studer, R. *Ontology-Oriented Information Extraction and Integration*. *Workshop on Language Technologies in Information and Knowledge Management (1999)*, 7th Conference on Computational Linguistics of the German Society for Language Technologies. Saarbrücken, Germany, October 7-8.
- Medline, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>
- Rindflesch, T., Rajan, J. and Hunter, L. (2000) Extracting Molecular binding relationships from biomedical text. *Proc. of the ANLP-NAACL, ACL*, pp. 188-195.
- Seeger, R. and Krebs E.G. (1995) *The MAPK Signaling Cascade*. *FASEB J*, 9(9), 726-735
- Sekimizu, T., Park, H.S., and Tsujii, J. (1998), Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Inf. Serv.* pp. 62-71.
- Stapley, B.J. and Benoit, G. (2000), *Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts*. *Pac. Symp. Biocomp.*, pp. 529-540.
- Thomas, J., Milward, D., and Ouzounis, C. (2000) Automatic extraction of protein interaction from scientific abstracts. *Pac Symp Biocomp.* pp. 384-395.
- Todirascu, A., Romary 'L., Bekhouche, D. (2002) *Vulcain - An Ontology-Based Information Extraction System*. *Natural Language Processing and Information Systems: 6th International Conference on Applications of Natural Language to Information Systems, NLDB 2002*, Stockholm, Sweden, 64-75.

A Web Service for Biomedical Term Look-Up

Ian Roberts, Henk Harkema, Rob Gaizauskas, Mark Hepple
initial.surname@dcs.shef.ac.uk

Department of Computer Science, University of Sheffield, UK

1 Introduction

Recent years have seen a huge increase in the amount of biomedical information that has become available in electronic format. Some of this information resides in structured databases and ontologies such as OMIM ([12]) and GO ([7]). On-line biomedical literature collections, e.g., MEDLINE,¹ form another important, unstructured source of biomedical information. As Hirschman et al. ([10]) point out, since names of biological entities such as genes and proteins provide the critical links across the different sources of information, automatic identification of these terms is an essential step in the process of managing the wealth of biomedical information that is available electronically.

In order to support term recognition in the biomedical domain, we have developed Termino, a large-scale terminological resource for text processing applications ([8], [9]). Since term recognition is of paramount importance to biomedical information processing, most, if not all applications in this domain require access to terminological knowledge. For this reason we are making Termino available to the biomedical text mining community as a web service. This service provides public, standardized access to Termino over the web, allowing integration of this resource into applications as a remotely located component. The web service delivers lexical look-up functionality: given a text, the service will return a version of the text in which occurrences of terms are identified and marked up with information from Termino. We have also implemented a web browser-based interface to the web service, which provides potential users with a simple way of exploring the utility of the web service.

In the remainder of this paper we will give a brief description of Termino and outline how we have implemented web service access to this resource.

2 Architecture and Functionality of Termino

Termino is a large-scale terminological resource for text processing applications. It includes a flexible, extensible relational database which stores large numbers of terms together with complex, heterogeneous information about these terms, including information of a morpho-syntactic nature, such as part of speech and morphological class; information of a semantic nature, such as quasi-logical form and links to concepts in ontologies; and provenance information, such as the sources of the information in the database. The design of the database also allows for links to connect synonyms and morphological and orthographic variants to one another, and to connect abbreviations and acronyms to their full forms. To ensure fast term look-up with Termino's potentially vast terminological database, the system comes equipped with a compiler for generating finite state machines from the strings in the terminological database. To recognize terms in text, the recognizer is run starting at each position, i.e., token, in the text. It will determine very quickly whether there are any strings in the database beginning at this point, and, if so, what information in the database these strings are associated with. This set-up turns Termino into a general terminological resource which is not restricted to any single domain or application. The database can be loaded with terms from multiple domains and compilation can be restricted to particular subsets of strings in the database by

¹<http://www.ncbi.nlm.nih.gov/PubMed/>

selection based on, for example, their source or other characteristics. In this way one can produce term recognizers that are tailored towards specific domains or specific applications within domains.

The contents of Termino's database are imported from existing, outside knowledge sources, such as the HUGO Gene Nomenclature database ([14]) and the Metathesaurus of the Unified Medical Language System (UMLS, [11]). Contents can also be induced from text corpora, e.g., MEDLINE citations. Termino thus enables uniform access to terminological information aggregated across many sources, without the need for multiple, source-specific terminological components within a text processing system. Term look-up in Termino provides immediate entry points into a variety of outside ontologies and other knowledge sources, making the information in these sources available to processing steps subsequent to term recognition. For example, using a recognizer compiled to include terms from the HUGO and OMIM databases, Termino will return the HUGO and OMIM identifiers for the gene names it recognizes in a text. These identifiers give access to the information stored in these databases about the gene, including alternative names, gene map locus, related disorders, and references to relevant papers.

Termino is designed to be the first component in a multi-component term processing system. Thus, term look-up as performed by Termino is not the end point of term processing. Look-up might return multiple matching terms for a given string, or for overlapping strings, and subsequent processes may apply to filter these alternatives down to the single option that seems most likely to be correct in the given context. Furthermore, more flexible processes of term recognition might apply over the results of look-up. For example, a term *grammar* can be provided for a given domain, allowing longer terms to be built from shorter terms that have been identified by term look-up.

For further details about the design and implementation of Termino the reader is referred to [8] and [9].

3 Term Look-up Web Service

Since term recognition is an important aspect of biomedical text processing, we think it would be useful to share Termino with the wider biomedical text mining research community. We have decided to do this by making Termino available in the form of a web service.

In the following sections we will give a brief overview of web services in general and point to a few natural language processing web service applications, describe the functionality of Termino as a web service, and discuss some issues concerning the implementation of this particular web service.

3.1 Web Services

A web service (e.g., [4], [15]) provides access to data or processing resources on the web through standard Internet protocols. Typically, using a web service proceeds according to the following scenario. First, a client looking for a particular service will consult a global registry of web services to find a host providing the desired service. The registry describes, in a standard manner, the functionality of a given web service and the way in which the web service can be accessed. Next, based on the information found in the registry, the client will select a particular web service and use it. Using a web service involves the transmission of standardized messages for invoking the service and receiving back the results of the invocation. The high degree of standardization of all aspects of the web service paradigm facilitates simple access across different software platforms to any resource packaged as a web service and enables easy integration of such a resource into distributed applications. Web services are thus an effective means for sharing resources between research groups. Their use promotes collaboration and prevents duplication of efforts spent on developing resources ([3]).

Further advantages of publishing Termino as a web service rather than releasing the resource as a downloadable software package include: users can access and integrate Termino into their applications without having to download and install the heterogeneous set of software components comprised in Termino; since the code and the data are located at our side, users do not have to download and install a new version of Termino each time the code is updated or the terminological database is expanded; web service access to Termino allows us to monitor its usage, providing us with a measure of its actual utility.

There is growing interest in the use of web services for natural language processing applications. For example, Biemann et al. ([1]) discuss the deployment of web service technology in the domain of cor-

pus processing. Dalli et al. ([3]) describe a general web service-based architecture for language resources and present details about the implementation of two web service applications built according to this architecture. Curran ([2]) shows how web services can be used to provide interfaces to a high performance infrastructure for natural language processing. Quasthoff and Wolff ([13]) give an example of a web service for dictionary look-up and terminology extraction. Gaizauskas et al. ([5]) describe an application in which web services are used to incorporate text mining capabilities into a workflow environment for supporting scientific discovery in bioinformatics. If, as expected, web services will become the dominating paradigm for communication within distributed systems, more and more natural language processing resources will be made available as web services.

3.2 Functionality

The web service built on top of Termino has been designed to implement a lexical look-up service: given a text, the service returns a version of the text in which term occurrences are identified and marked up with information from the Termino database. A request to the web service includes the text to be processed or its URL, so that the web service can download the text. Furthermore, in the service request the user can indicate the classes of terms that are to be tagged in the text, e.g., 'gene', 'protein', 'body part', etc. These classes are organized into a simple ontology, enabling the the user to choose a set of classes matching the semantic granularity of the intended application. The selection of term classes determines the set of recognizers compiled from Termino's database that will be run over the input text. In the current prototype, we offer a set of pre-compiled recognizers from which the user can make a choice, covering the term classes 'gene', 'disease or syndrome', and 'human protein'. The terms in the first class are imported from the HUGO and OMIM databases, the terms in the second class come from UMLS, and the terms in the third class originate from EBI's GOA project.² The web At a later stage, we may consider a scenario in which recognizers will be compiled at request time so that users can dynamically ask for mark-up of arbitrary term classes.

The response to a request to the web service is a text in XML format in which occurrences of terms are labeled with the term classes they belong to and are annotated with additional information from Termino's database. As Termino provides a lexical look-up service rather than full term identification and classification, terms that are assigned multiple classes are not disambiguated. Additional information for a term may include, for example, a UMLS unique concept identifier, HUGO and OMIM database identifiers, and assignments to nodes in the Gene Ontology. In the current implementation of the web service all information in the Termino database found for a term is returned; in the next version the user will be able to ask for only specific kinds of information from the database. The next version of the web service will also draw on the synonymy information stored in Termino to supply synonymy class identifiers for terms. These identifiers may be used to determine which terms in a text are synonymous.

It is possible that a text will contain terms with overlapping spans. For example, in the phrase *middle ear infection*, the sequence *middle ear* may be recognized as a 'body part, organ, or organ component' and the sequence *ear infection* may be marked-up as belonging to the class 'disease or syndrome'. Such overlapping terms cannot be represented using standard in-line XML mark-up. One solution to this problem is to select only one of the overlapping terms for mark-up. However, the decision about which term to keep and which term to discard is generally application-dependent and should therefore not be made by the term look-up service. We address this problem by adopting a slightly more complex XML encoding, which combines characteristics of in-line and stand-off mark-up, and which can represent overlapping terms. This approach allows users to choose amongst overlapping term possibilities for themselves.

Figure 1 shows the XML document returned by the Termino web service when the recognizers for the term classes 'gene' and 'disease or syndrome' are turned on, given an input document containing the single sentence *Since mutations in the gamma-crystallin encoding CRYG genes have previously been demonstrated to be the most frequent reason for isolated congenital cataracts, all 4 active CRYG genes have been sequenced.* The `Text` element contains the text of the input document into which `Node` elements have been inserted. These nodes are referenced in the `Annotations` element, which specifies the terms found in the text and their annotations. We see, for example, that the text between nodes 4 and 5, i.e., the

²<http://www.ebi.ac.uk/GOA/>

```

<?xml version="1.0" encoding="UTF-8" ?>
<TaggedDocument>
<Text>Since mutations in the gamma-crystallin encoding
<Node id="4"/>CRYG<Node id="5"/> genes have previously been
demonstrated to be the most frequent reason for isolated
congenital <Node id="0"/>cataracts<Node id="1"/>, all 4 active
<Node id="2"/>CRYG<Node id="3"/> genes have been sequenced.
</Text>
<Annotations>
  <Gene startNode="2" endNode="3">
    <omim_number>123730</omim_number>
  </Gene>
  <Gene startNode="4" endNode="5">
    <hgnc_id>2417</hgnc_id>
  </Gene>
  <Gene startNode="4" endNode="5">
    <omim_number>123730</omim_number>
  </Gene>
  <Gene startNode="2" endNode="3">
    <hgnc_id>2417</hgnc_id>
  </Gene>
  <Disease startNode="0" endNode="1">
    <cui>C0007388</cui>
  </Disease>
</Annotations>
</TaggedDocument>

```

Figure 1: Marked-up document

first occurrence of *CRYG*, belongs to the term class ‘gene’ and that this gene has OMIM number 123730 and HUGO identifier (hgnc.id) 2417. It is easy to see how this node-based scheme can be used to annotate overlapping terms. For example, the phrase *middle ear infection* will have nodes inserted between all adjacent tokens:

```

<Node id="0"/>middle <Node id="1"/>ear <Node id="2"/>infection <Node id="3"/>

```

The set of annotations for this phrase will contain the following elements:³

```

<Body_part startNode="0" endNode="2">
  <cui>C0013455</cui>
</Body_part>
<Disease startNode="0" endNode="3">
  <cui>C0029882</cui>
</Disease>
<Body_part startNode="1" endNode="2">
  <cui>C0013443</cui>
</Body_part>
<Disease startNode="1" endNode="3">
  <cui>C0699744</cui>
</Disease>
<Disease startNode="2" endNode="3">
  <cui>C0021311</cui>
</Disease>

```

The web service described above is complemented with a web-based GUI. The purpose of the GUI is to give the user an opportunity to assess the functionality of the web service without having to set up a web service client and engage in a full web service interaction. Through the GUI the user can submit a short

³A CUI is a unique concept identifier from UMLS.

text fragment for term look-up and inspect the results. As in the full web service, the user can select the classes of the terms which are to be looked up. The results of term look-up are presented in the format shown in figure 2.

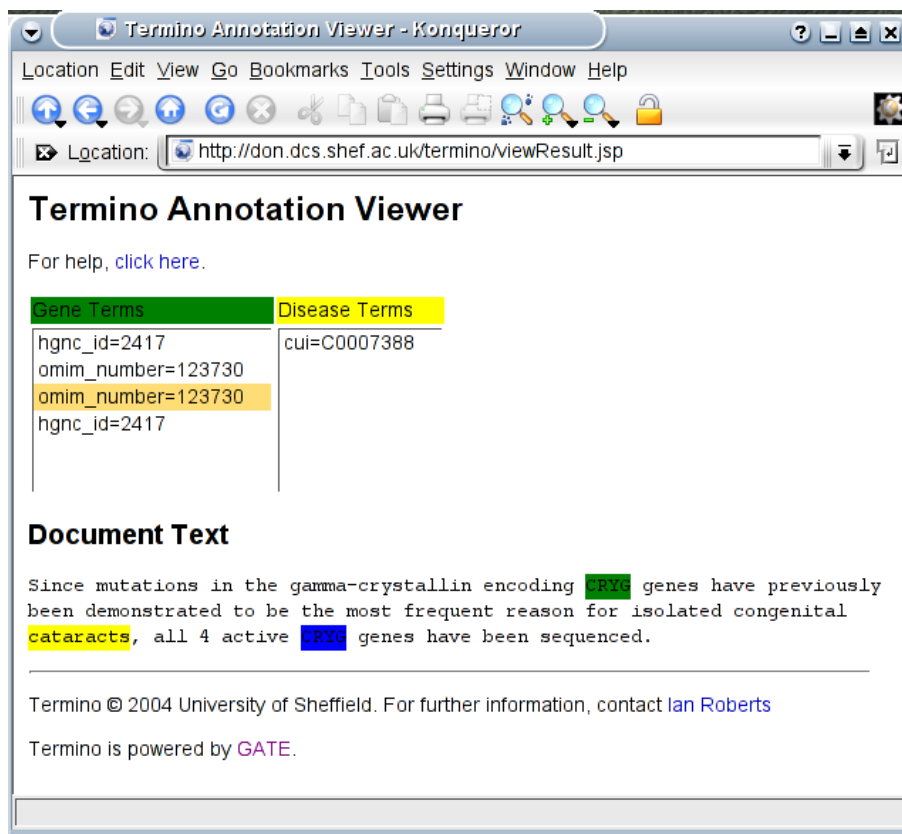


Figure 2: Screen shot of the web-based GUI

In the tables appearing above the document text each term class is assigned a color. The terms occurring in the text are marked-up in the colors of the term classes to which they belong. The annotation viewer also displays any additional information that is associated with a term. Clicking on a line in the table of a particular term class will highlight the term occurrence in the text to which the information given in that line applies. Vice versa, clicking on a term occurrence in the text will highlight the lines in the term class tables containing the information associated with this term.

3.3 Implementation

Termino is implemented as a collection of processing modules which run within the GATE architecture.⁴ The web service interface is implemented in Java, using the Apache Axis web services toolkit,⁵ running in the Apache Tomcat web server.⁶ In order to get a first impression of the performance of the web service, we submitted 100 MEDLINE abstracts to the web service for processing. The average size of these abstracts was 1.1 kB. Processing the abstracts took 1 minute and 52 seconds, i.e., approximately 1.1 second per abstract. On average, 11.7 terms were marked-up in each abstract.

The browser-based interface is implemented as a Java servlet which makes use of the same processing modules as the web service to annotate the supplied text. The results are rendered using JavaServer

⁴<http://www.gate.ac.uk/>

⁵<http://ws.apache.org/axis/>

⁶<http://jakarta.apache.org/tomcat/>

Pages (JSP) to generate an HTML page which uses a library of JavaScript functions to allow the user to explore the terms found by Termino. The web-based service can be found (at the time of writing) at <http://don.dcs.shef.ac.uk/termino/>, along with a WSDL definition for the web service interface.

4 Conclusion

Termino is a large-scale terminological resource for biomedical text processing, developed to provide term recognition capabilities for information extraction, retrieval, and navigation. It has been integrated into AMBIT, our platform for biomedical language processing ([6]). Since access to Termino could be helpful to other research groups wanting to integrate a terminological resource into their applications, we have made Termino available as a web service. A web service provides public, standardized access to a resource over the web. The Termino web service delivers lexical look-up functionality: given a text, the service will return a version of the text in which term occurrences are identified and marked up with information from the Termino database.

References

- [1] C. Biemann, S. Bordag, U. Quasthoff, and C. Wolff. 2004. Web Services for Language Resources and Language Technology Applications. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal.
- [2] J.R. Curran. 2003. Blueprint for a High Performance NLP Infrastructure. In: *Proceedings of the HLT/NAACL Workshop on Software Engineering and Architecture of Language Technology Systems*. Edmonton, Canada.
- [3] A. Dalli, V. Tablan, K. Bontcheva, Y. Wilks, D. Broeder, H. Brugman, and P. Wittenburg. 2004. Web Services Architecture for Language Resources. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal.
- [4] C. Ferris and J. Farrell. 2003. What are Web Services? In: *Communications of the ACM*, 46(6): 31.
- [5] R. Gaizauskas, N. Davis, G. Demetriou, Y. Guo, and I. Roberts. Forthcoming. Integrating Biomedical Text Mining Services into a Distributed Workflow Environment. In: *Proceedings of the UK e-Science All Hands Meeting 2004*. Nottingham, UK.
- [6] R. Gaizauskas, M. Hepple, N. Davis, Y. Guo, H. Harkema, A. Roberts, and I. Roberts. 2003. AMBIT: Acquiring Medical and Biological Information from Text. In: *Proceedings of the UK e-Science All Hands Meeting 2003*, S. Cox (ed.). Nottingham, UK.
- [7] The Gene Ontology Consortium. 2001. Creating the Gene Ontology Resource: Design and Implementation. In: *Genome Research*, 11(8): 1425-1433.
- [8] H. Harkema, R. Gaizauskas, M. Hepple, A. Roberts, I. Roberts, N. Davis, and Y. Guo. 2004. A Large Scale Terminology Resource for Biomedical Text Processing. In: *Proceedings of the NAACL/HLT Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*. Boston, USA.
- [9] H. Harkema, R. Gaizauskas, M. Hepple, N. Davis, Y. Guo, A. Roberts, and I. Roberts. 2004. A Large-Scale Resource for Storing and Recognizing Technical Terminology. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal.
- [10] L. Hirschman, A. Morgan, and A.S. Yeh. 2002. Rutabaga by Any Other Name: Extracting Biological Names. In: *Journal of Biomedical Informatics*, 35: 247-259.
- [11] L. Humphreys, D.A.B. Lindberg, H.M. Schoolman, and G.O. Barnett. 1998. The Unified Medical Language System: An Informatics Research Collaboration. In: *Journal of the American Medical Informatics Association*, 1(5): 1-13. <http://www.nlm.nih.gov/research/umls/>.
- [12] *Online Mendelian Inheritance in Man, OMIM (TM)*. 2000. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). <http://www.ncbi.nlm.nih.gov/omim/>.
- [13] U. Quasthoff and C. Wolff. 2003. Web Services in Language Technology and Terminology Management. In: *Proceedings of the 6th Terminology in Advanced Management Applications Conference*. Pretoria, South Africa.
- [14] H.M. Wain, M. Lush, F. Ducluzeau, and S. Povey. 2002. Genew: The Human Nomenclature Database. In: *Nucleic Acids Research*, 30(1): 169-171. <http://www.gene.ucl.ac.uk/nomenclature/>.
- [15] World Wide Web Consortium (W3C). 2004. *Web Services Activity*, <http://www.w3.org/2002/ws/>.

Towards a Semantic Lexicon for Biological Language Processing

Karin Verspoor
Los Alamos National Laboratory
verspoor@lanl.gov

It is well understood that natural language processing (NLP) applications require sophisticated lexical resources to support their processing goals. In the biomedical domain, we are privileged to have access to extensive terminological resources in the form of controlled vocabularies and ontologies, which have been integrated into the framework of the National Library of Medicine's Unified Medical Language System's (UMLS) Metathesaurus. However, the existence of such terminological resources does not guarantee their utility for NLP. In particular, we have two core requirements for lexical resources for NLP in addition to the basic enumeration of important domain terms: representation of *morphosyntactic* information about those terms, specifically part of speech information and inflectional patterns to support parsing and lemma assignment, and representation of *semantic* information indicating general categorical information about terms, and significant relations between terms to support text understanding and inference (Hahn et al, 1999). Biomedical vocabularies by and large commonly leave out morphosyntactic information, and where they address semantic considerations, they often do so in an unprincipled manner, for instance by indicating a relation between two concepts without indicating the type of that relation.

But all is not lost. The UMLS knowledge sources include two additional resources which are relevant – the SPECIALIST lexicon, a lexicon addressing our morphosyntactic requirements, and the Semantic Network, a representation of core conceptual categories in the biomedical domain. The coverage of these two knowledge sources with respect to the full coverage of the Metathesaurus is, however, not entirely clear. Furthermore, when our goals are specifically to process biological text – and often more specifically, text in the molecular biology domain – it is difficult to say whether the coverage of these resources is meaningful. The utility of the UMLS knowledge sources for medical language processing (MLP) has been explored (Johnson, 1999; Friedman et al 2001); the time has now come to repeat these experiments with respect to biological language processing (BLP). To that end, this paper presents an analysis of the UMLS resources, specifically with an eye towards constructing lexical resources suitable for BLP. We follow the paradigm presented in Johnson (1999) for medical language, exploring overlap between the UMLS Metathesaurus and SPECIALIST lexicon to construct a morphosyntactic and semantically-specified lexicon, and then further explore the overlap with a relevant domain corpus for molecular biology.

The UMLS as a Lexical Knowledge Source

There have been several investigations of the UMLS as a lexical knowledge source. McCray et al (2001) evaluate the nature of strings in the UMLS Metathesaurus with respect to their likelihood of appearing in a natural language corpus. They found that only 10% of the strings in the Metathesaurus occurred in their MEDLINE corpus (representing one year of MEDLINE abstracts), but were able to identify some properties associated with the strings that could be used to filter out strings that are unlikely to occur naturally in a corpus. While the authors suggest that occurrence of a term in the Metathesaurus opens the possibility of accessing more extensive domain knowledge about that term, they do not explore the nature of that domain knowledge for the terms they find in their corpus, and do not explore the overlap of those terms with other UMLS resources.

Friedman et al (2001) quantitatively compare a lexicon developed manually for their MEDLEE system with a lexicon derived automatically from the UMLS, with respect to the task of processing clinical information in patient reports. They found the UMLS-derived lexicon to lead to poor performance relative to their own lexicon. The results do not, however, invalidate the UMLS as an important source of lexical information, as they may be a reflection of the completeness of the existing MEDLEE lexicon for the task evaluated. The authors argue that using the UMLS can substantially reduce the manual effort in constructing a lexicon.

Johnson (1999) explores the construction of a lexical resource from the UMLS in support of processing of medical narrative, specifically utilizing a corpus of discharge summaries from hospital visits. Johnson explores the overlap between the Metathesaurus, the SPECIALIST lexicon, and a domain corpus, and presents some strategies for handling semantic ambiguities that arise during the mapping of terms in the different UMLS resources. Johnson found that while 79% of the distinct lexical forms in his corpus occurred in the SPECIALIST lexicon, only 38% of those forms occurred in the semantic lexicon of more than 75,000 entries derived from intersecting the Metathesaurus and the SPECIALIST lexicon – so, only 38% of terms in the corpus could be expected to have both morphosyntactic and semantic information derived from the UMLS. Johnson points out this may reflect the fact that the Metathesaurus may contain many complex medical terms that should not be considered lexical items, and that furthermore may successfully be incorporated into the lexicon by assuming that they are nouns.

Methods

We follow Johnson (1999) and explore the overlap in the UMLS Metathesaurus and the SPECIALIST lexicon to establish a baseline semantic lexicon, and then investigate its relevance for a corpus in the molecular biology domain. We utilize the 2003AC UMLS release. As our domain corpus, we utilize 28,874 full text articles from the Journal of Biological Chemistry (JBC), spanning the years 1998-2002, originally obtained for the 2003 BioCreAtIvE competition. While we realize that this is not a sample representative of the full domain of molecular biology, it is representative of a significant portion of that domain, and the results on JBC texts should be indicative of the coverage of our semantic lexicon for this domain. We felt it preferable to use a corpus of full text articles rather than a corpus of abstracts derived from MEDLINE in order to more completely assess coverage of the relevant language.

The steps for building and evaluating our semantic lexicon are as follows:

- Lexemes in the SPECIALIST lexicon are matched to terms in the Metathesaurus. We load in all the strings represented in the SPECIALIST LRAGR file, and attempt to match Metathesaurus strings extracted from the MRCON file to these strings. This is done by considering different kinds of matches:
 - Exact match
 - Match after uppercasing the first letter of the SPECIALIST string
 - Match after uppercasing the first letter of each word of the SPECIALIST string
 - Match after uppercasing the entire SPECIALIST string
 - Other case insensitive match
 - Match (any of the above types) after stripping the Metathesaurus string of “, NOS” or “<1>”, “<2>”, etc. at the end
 - Finally, consider whether each of the constituent tokens of a multi-token (space containing) Metathesaurus string occurs in the SPECIALIST lexicon (after removal of words consisting of all numbers or punctuation), in order to assume a compositional analysis of the term
- Filter the resulting lexicon (a subset of the original SPECIALIST lexicon tied to specific Metathesaurus terms) by removing any terms for which the corresponding Metathesaurus string is not associated with a semantic type through one of its associated concepts. There may be concepts for which the UMLS does not provide semantic information, and therefore they do not satisfy our lexical constraints requiring both morphosyntactic and semantic information.
- Search the domain corpus for occurrences of any lexical variant of each term in our semantic lexicon (obtaining lexical variants from the UMLS lexical tools), and track any matches in order to establish the coverage of the lexicon.

Exact matches	58,918	3.0%
First letter uppercase	67,765	3.5%
First letter, all words uppercase	13,922	0.7%
Entire string uppercase	12,961	0.7%
Other case insensitive match	1,982	0.1%
Stripped term matches	5,945	0.3%
Total direct matches	161,493	8.2%
Constituent matches	1,548,389	79.0%
Total matches	1,709,882	87.3%

Table 1: Matches between the UMLS Metathesaurus terms and the SPECIALIST lexicon

Results

Our results on matching between the SPECIALIST lexicon and the Metathesaurus, shown in Table 1, indicate that the proportion of Metathesaurus terms directly occurring (through some matching paradigm) in the SPECIALIST lexicon is in fact slightly less than Johnson's (1999) finding of 12% at 8.2%. This is due to the incredible growth in the Metathesaurus in the past few years; Johnson reports finding 630,658 unique strings in the Metathesaurus, while the version we worked with contains 1,959,516 unique strings. The SPECIALIST lexicon has grown as well (from 164,850 distinct lexical forms to 292,979), but clearly not at pace with the Metathesaurus. This result is in line with Johnson's observation that many of the terms in the Metathesaurus are probably not appropriate for recording directly in the SPECIALIST lexicon. However, upon inspection of the constituent structure of Metathesaurus terms, we found that for a large proportion of terms (79%), each of the constituent members of the (multi-word) term could be found in the SPECIALIST lexicon. This opens the possibility of a compositional analysis for many Metathesaurus terms, though it doesn't address the assignment of semantic type to the term as a whole.

The number of unique SPECIALIST terms matched by Metathesaurus terms was 108,295. These string matches were used to create a lexicon containing 96,205 unique entries from the SPECIALIST lexicon (where a given term may correspond to multiple lexical entries due to the morphosyntactic ambiguity of the term, and a given lexical entry may correspond to multiple terms due to lexical variation) by identifying each of the lexical entries a matched string may correspond to. This is 52% of the complete SPECIALIST lexicon (of 183,301 entries). Filtering this lexicon according to the constraint of having a semantic type for each had no impact whatsoever – we found that each of the 78,595 unique Metathesaurus concepts matched to a SPECIALIST lexicon term¹ was also associated with a semantic type in the Metathesaurus, so there was no reduction in the lexicon.

We next explored the overlap of the resulting lexicon with our domain corpus, by looking for matches between tokens in the corpus and any lexical form associated with the 96,205 entries in our subset of the SPECIALIST lexicon (whether or not that exact form occurred in the initial Metathesaurus term set). We split each of the 28,874 JBC files into tokens after stripping HTML tags and converting HTML character entities. We investigated several different ways in which a token could match a lexical entry:

- Exact single token match: the token occurs in the lexicon exactly as it appears in the text
- Case-insensitive single token match: the token in the text matches a lexical entry when matched case insensitively
- Exact multi-token term match: the token starts a phrase in the text that exactly matches a multi-token term in the lexicon
- Case-insensitive muti-token term match: the token starts a phrase in the text that matches a multi-token term in the lexicon when matched case insensitively

¹ Note that this number is significantly lower than the number of Metathesaurus term matches reported. This is because several distinct terms in the Metathesaurus may correspond to the same concept.

	count	% of base set
Total number of files processed	28,874	
Basic token matches		
Number of tokens	156,608,748	
Single token matches	121,552,230	77.6%
Additional matches with case insensitivity	9,419,429	6.0%
Multi-token term matches	2,866,226	1.8%
Additional multi-token term matches with case insensitivity	261,128	0.2%
Number of unique tokens	1,898,320	
Unique unmatched tokens	1,836,148	97%
Unique unmatched numeric tokens	78,770	0.5%
Matches for tokens following hyphenation relaxation		
Number of tokens relaxed	6,869,993	
Relaxed tokens directly matching multi-token term	157,529	2.3%
Tokens starting (longer) multi-token term match	14,100	0.2%
Additional matches with case insensitivity	0	0.0%
Number of constituent tokens	13,994,307	
Tokens with some constituent match	4,899,189	71.3%
Number of constituent tokens matching	7,396,976	52.9%
Lexicon matches		
Unique lexemes in lexicon	292,979	
Unique lexical entry IDs in lexicon	96,205	
Unique single token lexemes in lexicon	268,617	
Unique multi-token lexemes in lexicon	24,362	
Unique single token lexemes matched	62,172	23.1%
Unique multi-token lexemes matched	15,290	62.8%
Unique lexical entry IDs matched	59,199	61.5%

Table 2: Matches between the derived lexicon and the domain corpus

- “Relaxed” hyphenated token match: for single tokens containing hyphens that did not match a lexical entry in some way, we generated a variant of the token with the hyphens replaced by spaces, effectively generating a multi-token term out of the original single token. The following matches were then attempted:
 - Match (exact) of the relaxed token string to a multi-token term in the lexicon
 - Matching (exact and case insensitive) where the relaxed token string starts a phrase in the text matching a (longer) multi-token term in the lexicon
 - If the relaxed token string did not match a multi-token term in the lexicon, attempt to match each of the constituent words of the string to a lexeme

The results appear in Table 2. We see that over 77% of the tokens in the corpus match exactly to a lexeme in the lexicon, with a total of 83% matching when case insensitive matches are allowed. Though this corresponds to only approximately 3% of the distinct tokens found in the corpus, the high coverage of the corpus as a whole indicates that this 3% corresponds to the most frequent tokens in the corpus. The lexicon includes the main content-bearing terms of the domain, in addition to the expected grammatical function words such as “and”, “the”, etc.

Inspection of the tokens which did not match any lexeme in the lexicon show that a large proportion of the unmatched tokens are numeric tokens. This accounts for an additional 7,969,674 of the tokens (5%), though they correspond to only 0.5% of all distinct tokens. Other frequently unmatched token types correspond to chemical formulas (e.g. “K+”), gene/protein names (e.g. “ERK2”), typographical errors (e.g. “negaitve”), protein sequences (e.g. “CACAGAGGATGGGTAACTCCAG”), proper names, some tokens that only occur as part of a multi-token term (e.g. “de” in “de novo” or “vitro” in “in vitro”), as well as many that seem to derive from errors in our tokenization or problems with handling of UNICODE characters. Many of these could be handled by specific tokenization and token tagging strategies, rather than requiring that the terms be enumerated in the lexicon.

The lexicon does contain a significant number of terms which were not found in the corpus, since only 62% of the lexical entries had a match (on at least one of its lexical variants) in the corpus, but it does not necessarily follow that the remaining 38% of the lexicon is irrelevant for biological language processing, as it could be that our corpus is not fully representative of the domain.

Conclusions

We have found sufficient overlap with our derived semantic lexicon to justify the use of the UMLS resources as a starting point for a lexicon for Biological Language Processing, on the basis of lexical overlap between a lexicon derived from a combination of the UMLS Metathesaurus and the SPECIALIST lexicon, and the terms in a domain corpus. Over 77% of the tokens in the domain corpus are found (through exact match) in the derived lexicon, though only 3% of the unique tokens in the corpus are covered. This shows that the terms captured in the derived lexicon cover the most frequent, and likely the most content-bearing, terms in the domain corpus. Through augmentation with some domain-specific tokenization and named entity extraction, this lexicon can be extremely valuable for BLP.

There remain questions about the utility of the UMLS Semantic Network for BLP. Although we have established a core lexicon for which we have the basic required lexical information – morphosyntactic and semantic information – we have not investigated any potential shortcomings of the UMLS Semantic Network. There are 135 semantic types and 54 relationship types represented in the 2003AC version of the Semantic Network; the number of types is quite small given the complexity of the biomedical domain, and this begs the question of whether it adequately characterizes the semantic distinctions needed for BLP. In contrast, the Gene Ontology resource (Ashburner et al, 2000) contains over 16,000 concepts grouped hierarchically and therefore in principle represents a much more fine-grained semantic breakdown of the domain. The GENIA ontology under development (Ohta et al, 2002) is focused on cell signaling reactions in humans and as such characterizes concepts specific to those processes, again likely to be much more fine-grained than the broad UMLS ontology. The relative utility of different ontologies should be investigated.

Furthermore, we have not explored whether the semantic types associated with terms in our derived lexicon are correct for the specific usages found in our corpus. This is an important issue to be investigated, as a semantic lexicon is only useful to the extent that it captures the appropriate semantics. Finally, we have not assessed the impact of semantic ambiguity on our lexicon – how many of the lexical items are multiply ambiguous, how much of this ambiguity is appropriate to the biological language, and how can we best deal with this ambiguity? These are the questions that we must answer to fully assess the utility of a UMLS-based lexicon for biological language processing.

References

- Ashburner, M; Ball, C.A.; and Blake, J.A. et al [2000]. "Gene Ontology: Tool for the Unification of Biology", *Nature Genetics*, v. 25:1, pp 25-29.
- Friedman, Carol, Hongfang Liu, Lyuda Shagina, Stephen Johnson, George Hripcsak [2001]. Evaluating the UMLS as a Source of Lexical Knowledge for Medical Language Processing. *Proc. AMIA 2001*; 189-193.
- Hahn, Udo, Martin Romacker, Stefan Schulz [1999]. How Knowledge Drives Understanding – Matching Medical Ontologies with the Needs of Medical Language Processing. *Artificial Intelligence in Medicine*, 15:25-51.
- Johnson, Stephen B [1999]. A Semantic Lexicon for Medical Language Processing. *Journal of the American Medical Informatics Association*, 6:3, 205-218.
- McCray, Alexa T., Olivier Bodenreider, James D. Malley, Allen C. Browne [2001]. Evaluating UMLS Strings for Natural Language Processing. In the *Proceedings of the AMIA Annual Symposium 2001*; 448-452.
- Ohta, Tomoko, Yuka Tateisi, Jin-Dong Kim [2002]. GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In the *Proceedings of the Human Language Technology Conference (HLT 2002)*.

Abstracts for Invited Talks

Report on the BioCreAtIvE Workshop Granada, 2004

Christian Blaschke¹, Lynette Hirschman²,
Alexander Yeh², Alfonso Valencia¹

¹Centro Nacional de Biotecnología, Universidad Autónoma, Madrid, Spain

²The MITRE Corporation, Bedford, MA

Abstract

The first BioCreAtIvE Workshop (Critical Assessment of Information Extraction in Biology) was held in Granada, Spain March 28-31, 2004. The goal on the workshop was to provide a set of common challenge evaluation tasks to assess the state of the art for text mining applied to biological problems. The assessment focused on two tasks. The first dealt with extraction of gene or protein names from text, and their mapping into standardized gene identifiers for three model organism databases (fly, mouse, yeast). The second task addressed issues of functional annotation, requiring systems to provide gene ontology annotations for proteins, given full text articles. Overall, 27 groups participated in the assessment, including 18 for gene/protein name extraction, and 9 groups for the GO functional annotation task.

The results for gene/protein name extraction showed that a number of groups (4) were able to extract general gene names from sentences of MEDLINE abstracts at over 80% balanced precision and recall. For the name normalization subtask, the results ranged from a high for yeast of 92% balanced precision and recall, to somewhat lower scores for fly (82%) and mouse (79%), due to extensive ambiguity among gene synonyms and overlap with standard English vocabulary.

For the functional annotation task, systems were asked to identify a segment of text as evidence for a GO annotation, given the protein. The annotation and the text were reviewed by expert GOA annotators at EBI for validity. When both protein name and the GO annotation were given, several systems provided correct evidence for the GO predictions 25-30% of the time; two systems provided a much higher rate of correct predictions (50% and 75-80%) by predicting only for high confidence cases. When the systems were given only the protein name, the results were significantly lower (~10% for systems providing predictions for all proteins and ~30-35% for the high precision systems providing only a few answers).

Enhancing Access to the Bibliome: The TREC Genomics Track

William R. Hersh
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
3181 SW Sam Jackson Park Rd., BICC
Portland, OR, USA 97239
hersh@ohsu.edu

The growing amount of scientific research in genomics and related biomedical disciplines has led to a corresponding growth in the amount of on-line data and information, including scientific literature. A challenge for biomedical researchers is how to access and manage this ever-increasing quantity of information.

The Text Retrieval Conference (TREC) is an annual activity of the information retrieval (IR) research community sponsored by the National Institute for Standards and Technology (NIST). TREC aims to provide a forum for evaluation of IR systems and users. Activity is organized into “tracks” of common interest, such as question-answering, multi-lingual IR, Web searching, interactive retrieval, and, starting in 2003, IR in the genomics domain.

The first year of the TREC Genomics Track (2003) was very successful, with a total of 29 groups participating. There were two tasks for the 2003 track. The first was an ad hoc retrieval task using MEDLINE records (titles, abstracts, and human-assigned MeSH terms) as documents, gene names (and their synonyms) as queries, and Gene Reference into Function designations (GeneRIFs) as relevance judgments. The second task was to nominate the annotation text of the GeneRIF given the MEDLINE record and full text of the article.

The second year of the track will also have two tasks. An ad hoc retrieval task will use topic statements derived from real biologists’ information needs, a 10-year subset of MEDLINE, and relevance judgments done in the usual TREC manner. A categorization task will have participants detect the presence or absence of experimental evidence warranting GO code assignment for mouse genes in the full text of journal articles.

The track is sustained with a National Science Foundation Information Technology Research grant that provides funding through 2008. Background on the motivation and evolution of the track can be found on the track Web site (<http://medir.ohsu.edu/~genomics/>). The Web site also contains an overview paper from the 2003 track as well as the protocol for the 2004 track.

BioMinT : A Database Curator's Assistant for Biomedical Text Processing

Anne-Lise Veuthey

Teresa K. Attwood¹, Paul Bradley¹, Walter Daelemans⁵, Luc Dehaspe², Frederique Durant⁵, Melanie Hilario⁶, Jee-Hyub Kim⁶, Alex L. Mitchell¹, Johann Petrak³, Violaine Pillet⁴, Alexander K. Seewald³, Kristof Van Belleghem², Anne-Lise Veuthey⁴, Marc Zehnder⁴

¹School of Biological Sciences, University of Manchester, Manchester, United Kingdom, ²PharmaDM, Leuven, Belgium, ³Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria, ⁴Swiss Institute of Bioinformatics, Geneva, Switzerland, ⁵University of Antwerp, Antwerp, Belgium, ⁶University of Geneva, Geneva, Switzerland

The high quality of many biological databases is guaranteed by their information content which is extracted and synthesized from the scientific literature by biological experts. Such a manual annotation procedure is time-consuming. Hence, information extraction methods are very promising in facilitating the process of literature screening.

The goal of the BioMinT project is to develop a generic text mining tool that assists database manual annotation by: (1) interpreting diverse types of query; (2) retrieving relevant documents from the biological literature; (3) extracting the required information, and (4) providing the result as a database slot filler or as a structured report. The development of the BioMinT system has followed a strictly problem-oriented approach. All decisions relative to prototype design have been based on requirements from those who will use the final product in their daily work, i.e. the curators of Swiss-Prot - the knowledgebase component of the UniProt resource (1) - and PRINTS - the protein family fingerprint database (2) -, as well as biological researchers.

The core of the system is composed of an information retrieval module consisting in a meta-query engine wrapped around the PubMed server. The followed strategy ensures a high recall of documents from Medline by expanding the query with related terms. For gene and protein names, such an expansion is done using a synonym database constructed from existing resources of model organisms. The retrieved documents are then filtered, categorized and ranked according to their relevance with regard to the query. The initial prototype implements simple indexing algorithms for this task. We plan to improve this step using methods based on semantic-related criteria. Interactivity is a main feature of the module: a user interface provides control over each step of the query process.

The second system's module, which deals with information extraction, is still under development. It is based on the integration of adaptive natural language processing (NLP) techniques, domain-specific knowledge, and relational and statistical data mining techniques. A first step consists in the customization of a memory-based shallow parser (3) to biological text, using a training procedure on the GENIA corpus (4). Then, diverse machine-learning methods are trained to extract relevant sentences using NLP-analyzed pre-annotated corpora, i.e. collections of documents in which specific fragments containing information on a given topic were carefully tagged by domain experts. The performances of the different learning methods are under evaluation by the biological experts.

(1) Rolf Apweiler, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh. UniProt: the Universal Protein knowledgebase *Nucl. Acids. Res.* 2004 **32**: D115-D119

(2) Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. & Zygouri, C. (2003). "PRINTS and its automatic supplement, prePRINTS." *Nucleic Acids Res.*, **31**(1), 400-402.

(3) Halteren, H. van, Zavrel J., Daelemans W. (2001) "Improving accuracy in word class tagging through combination of machine learning systems." *Computational Linguistics* **27** (2), 199-230.

(4) Ohta, Tomoko, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. (2002). GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In the *Proceedings of the Human Language Technology Conference (HLT 2002)*. pp73-77.

CASP: Critical Assessment of Techniques for Protein Structure Prediction

Anna Tramontano
University of Rome "La Sapienza"

Abstract

The CASP community wide experiment critically assesses the state-of-the-art in the prediction of protein structure from sequence and it has been conducted on a two year cycle for the last decade, beginning in 1994.

The primary goals are to establish the capabilities and limitations of current methods of modeling protein structure from sequence, to determine where progress is being made, to determine where the field is held back by specific bottlenecks, and to compare the results of automatic prediction servers with manually submitted predictions. Methods are assessed on the basis of the analysis of tens of thousands blind predictions of protein structure submitted by a large number of prediction teams from around the world.

Such objective testing is the only way to obtain a useful measure of the value of particular approaches to prediction, what the defects are, and most important, where effort may most effectively be focused to move the field forward. CASP provides a forum in which there is a thorough examination of the outcome of the predictions - what went right, what went wrong, and where possible, to provide an understanding of why. For members of the structural biology community not directly involved in structure prediction, the results provide a reasonable guide to the current state of the art. For the prediction community, the results provide a new and sharper sense of direction. Finally, we can begin to measure progress in the field over time.

Another major challenge in genomics is the discovery of function. This might be even more difficult than structure prediction, but it is time that we try to see whether there are methods that can help effectively in this area.

To this aim a new category was added to CASP: function prediction. There are several types of predictions that can be submitted for each target: GO category, post-translational modification type and location, predicted binding to other molecules, functional sites, etc.

The evaluation of methods in this case is even more challenging, as we have no way of knowing when the unknown function of a protein will be discovered and because it is quite difficult to give a complete dictionary of possible functional prediction. Hopefully, though, the analysis of the predictions that we are collecting right now will allow us to improve the suggested type of functional classification.

If some property of a protein is consistently predicted by many predictors, we will approach experimental groups to have it verified. In any case, if and when the function of the target proteins will become known we will assess the predictions and evaluate their accuracy. This is but a first step and it is still in an experimental phase, but we are confident that it will nevertheless provide valuable information to the scientific community.

EVA: Automatic System for the Evaluation of Structure Prediction Servers

Burkhard Rost
Columbia University

Abstract

EVA (<http://www.rostlab.org/eva/>) is a web server for evaluation of the accuracy of automated protein structure prediction methods. The evaluation is updated automatically each week, to cope with the large number of existing prediction servers and the constant changes in the prediction methods. EVA currently assesses servers for secondary structure prediction, contact prediction, comparative protein structure modelling, and threading/fold recognition. Every day, sequences of newly available protein structures in Protein Data Bank are sent to the servers and their predictions are collected. The predictions are then compared to the experimental structures once a week; the results are published on the EVA web pages. Over time, EVA has accumulated prediction results for a large number of proteins, ranging from hundreds to thousands, depending on the prediction method. This large sample assures that methods are compared reliably. As a result, EVA provides useful information to developers as well as users of prediction methods. In the framework of EVA, we have analysed fold recognition, secondary structure and contact prediction servers from CAFASP3. We observed that the sequence-unique targets from CAFASP3/CASP5 were not fully representative for evaluating performance. For all three categories, we showed how careless ranking might be misleading. We compared methods from all categories to experts in secondary structure and contact prediction and homology modellers to fold recognisers. While the secondary structure experts clearly outperformed all others, the contact experts appeared to outperform only novel fold methods. Automatic evaluation servers are good at getting statistics right and at using these to discard misleading ranking schemes. We challenge that to let machines rule where they are best might be the best way for the community to enjoy the tremendous benefit of CASP as a unique opportunity for brainstorming.