Grounding Gene Mentions with Respect to Gene Database Identifiers

Ben Hachey, Huy Nguyen, Malvina Nissim, Bea Alex, Claire Grover

Institute for Communicating and Collaborative Systems
{bhachey|mnissim|v1balex|grover}@inf.ed.ac.uk
University of Edinburgh, United Kingdom

Department of Computer Science
htnguyen@stanford.edu
Stanford University, United States

Abstract

We describe our submission for task 1B of the BioCreAtIvE competition which is concerned with grounding gene mentions with respect to databases of organism gene identifiers. Several approaches to gene identification, lookup, and disambiguation are presented. Results are presented with two possible baseline systems and a discussion of the source of precision and recall errors as well as an estimate of precision and recall for an organism-specific tagger bootstrapped from gene synonym lists and the task 1B training data.

1. Introduction

We describe our submission for task 1B of the BioCreAtIvE competition.[1] Task 1B was concerned with grounding gene entities. Provided an organism database containing unique gene identifiers with lists of synonyms and an abstract, the system creates a list of unique gene identifiers for genes that are mentioned in the abstract, including explicit mentions as well as those implicit in mentions of gene mutants, alleles, and products. The task was evaluated for three target organisms: fly (Drosophila melanogaster), yeast (Saccharomyces cerevisiae), and mouse (Mus musculus). The unique identifier returned must come from the appropriate organism.

Consider the following abstract snippet from the fly evaluation data (the gene entities are highlighted in bold):

> Since **Dpp** and **Gbb** levels are not detectably higher in the early phases of cross vein development, other factors apparently account for this localized activity. Our evidence suggests that the product of the **crossveinless 2 gene** is a novel member of the BMP-like signaling pathway required to potentiate **Gbb** of **Dpp** signaling in the cross veins. **crossveinless 2** is expressed at higher levels in the developing cross veins and is necessary for local BMP-like activity.

The system output should singly list the unique identifiers for the **Dpp**, **Gbb**, and **crossveinless 2** genes:

FBgn0000395	crossveinless 2
FBgn0000490	Dpp
FBgn0024234	Gbb

A synonym database was available for each of the target organisms. These list a number of synonyms for each unique gene identifier. Consider the following fly examples:

[1] http://www.mitre.org/public/biocreative/

ID	SYNONYMS
FBgn0000395	CG15671, CT35855, **crossveinless 2**, cv 2, cv-2
FBgn0000490	CG9885, DPP, DPP C, DPP-C, **Dpp**, Haplo insuffi - cient, Haplo-insuffi cient, Hin d: Haplo insuffi cient, Hin-d: Haplo-insuffi cient, M(2)23AB, M(2)LS1, Tegula, Tg: Tegula, blink, blk: blink, decapentaplegic, dpp, heldout, ho, ho: heldout, l(2)10638, l(2)22Fa, l(2)k17036, short-vein, shv
FBgn0001105	CG10545, G beta, G betab, G protein &bgr; subunit, G protein &bgr;-subunit 13F, G protein beta 13F, G protein beta subunit, G protein beta subunit 13F, G protein beta-subunit 13F, G&bgr;, G&bgr;13F, G-&bgr;b, G-beta, G-betab, G-protein &bgr; 13F, G-protein beta 13F, G¡down¿&bgr;¡/down¿ brain, Gb13F, **Gbb**, Gbeta, Gbeta brain, Gbeta13F, anon EST:Liang 1.22, anon-EST:Liang-1.22, clone 1.22, dg&bgr;, dgbeta
FBgn0017531	Spal**crossveinless 2**, Spal\crossveinless-2, **crossveinless 2**, crossveinless-2
FBgn0018552	Dpse\cv2, crossveinless, **crossveinless 2**, crossveinless-2, cv
FBgn0024200	CG9936 Pap/Trap, Scad78, Suppressor of constitutively activated **Dpp** signaling 78, TRAP240, bli, blind spot, bls, dTRAP240, flytrap, l(3)L7062, l(3)rK760, pap, pap/dTRAP240, poils aux pattes
FBgn0024234	60A, CG5562, **Gbb**, Gbb 60A, Gbb-60A, SixtyA, TGF&bgr;-60A, TGFbeta 60A, TGFbeta-60A, Tgf&bgr;-60A, Tgfb 60, Tgfb-60, Tgfbeta 60A, Tgfbeta-60A, Transforming growth factor &bgr; at 60A, Transforming growth factor beta at 60A, gbb, gbb 60A, gbb-60A, gcn, gcn: gonial cell neoplasm, gcn: gonial-cell-neoplasm, glass bottom boat, glass bottom boat 60A, glass bottom boat-60A, l(2)60A J, l(2)60A-J, tgfb 60A, tgfb-60A, vgr/60A
FBgn0044017	Scad67, Suppressor of constitutively activated **Dpp** signaling 67

One way of approaching the task is through the synonym lists. In this case a look-up or pattern matching method is used to see if any of the synonyms occur in an abstract. The difficulty here lies in distinguishing between matches that actually correspond to gene entities from false positives. This is particularly problematic with mouse and fly gene entities whose synonyms include such common English words as **with**, **at**, and **yellow**.

We focused on an alternative approach where the task is viewed as a named entity recognition problem. Named entity recognition (NER) can be viewed as consisting of three main steps. First, the boundaries of the entities within a text are determined. Second, the identified entities are classified. (This is not relevant as gene is the only entity type we are concerned with.) The first and second steps are addressed in task 1A (Finkel et al., 2004). Third, the entities

are grounded with respect to its denotation in the world (or model of the world). This is the focus of task 1B.

The remainder of the paper describes our approaches and results. Section 2. presents the high-level architecture of our system and different ways to approach the sub-tasks. Section 3. contains results and analysis. Finally, we briefly discuss related work in Section 4.

2. System Description

We approached task 1B as consisting of three primary sub-problems: identifying gene text in an abstract, matching the tagged gene text against the synonym list, and choosing a single unique identifier for ambiguous synonyms. In experiments with a variety of methods, we found different configurations for different organisms to be better than a unified approach. The following subsections discuss different approaches to the three sub-problems.

2.1. Gene Text Identification

Gene text fragments were extracted from the abstracts using the same Conditional Markov Model tagger developed for task 1A (Finkel et al., 2004). A tagger was trained for each organism using the organism-specific training data available in task 1B. Gene text fragments in the training abstracts were labeled automatically by identifying potential genes using regular expression fuzzy matching (described below), and filtering out matches that did not correspond to gene IDs listed in the gold standard gene ID list.

Since we only resolve gene text fragments which match phrases in the synonym lists, to improve the precision of the gene taggers, we mark the known gene text fragments explicitly by setting gazette features (whether the phrase containing the current word appeared in the synonym list). The tagger can then immediately reject any word that does not appear inside a known synonym, and then use other features to distinguish between valid and erroneous genes.

Starting with the current word and gazette, we attempted to add other features to improve performance. We obtained the best F1 performance on the development set using only the current word, gazette , and POS tag sequence features (t_{-1},t,t_{+1}; t_{-1},t; t,t_{+1}). In general, the word features seemed to be the strongest indicators that a gene text fragment was relevant. Expanding the feature set beyond these simple features degraded recall, most likely because the other features were present in both relevant and irrelevant fragments, tending to prefer the more frequent irrelevant hypothesis. To further improve recall, we lowered the threshold for classifying a word as a gene by a fixed amount determined empirically (since resolving an entity with an organism gene list filters output in a way that enhances precision).

2.2. Synonym List Lookup

We identified two main approaches to improving recall for matching tagged text against synonym entries: adding synonym resources and fuzzy matching. A preliminary search for additional synonym resources gave the impression that the synonym lists provided contain a large proportion of the edited and formalised material that is available. While an attempt could be made to expand synonym lists by mining MEDLINE abstracts, we focused our attention on improving recall through fuzzy matching through three main appraches.

Regular expressions. We converted each synonym list into a regular expression that would match some of the typical variations used in biomedical nomenclature. Regular expression rules were created to accommodate: case folding, optional dashes and other punctuation, optional spaces, British/American spelling variations (e.g. our/or), and substring matching. Matching was performed by running each of the regular expressions over the tagged gene text, and selecting the expression that matched the most text.

Edited lookup tables. Hash tables provide a highly efficient data structure for lexical lookup. However, hash tables consisting of only the synonyms as they appear in the provided lists gives very low recall, especially when using automatically tagged text as lookup keys. The baseline look-up table approach discussed in 3.2. yields F-scores of 22.8 to 56.6, 51.7 to 59.1, and 50.0 to 64.3 respectively for fly, mouse and yeast development test sets depending on the gene named entity recognition used. To try and improve recall using an efficient hash lookup, we created a number of normalising operations (Table 1). Edit operations are applied when the look-up table is being created and when a piece of tagged text is being checked against the table. The organisation of these operations was optimised per organism. Matches requiring fewer edits are preferred.

EDIT	DESCRIPTION	EXAMPLE
CAP	Capitalisation	ZFH-2 $\mapsto$ zfh-2
EX	Extra text	zfh-2 gene $\mapsto$ zfh-2
WR	Writing conventions	zfh-II $\mapsto$ zfh-2
PUNC	Punctuation	zfh-2 $\mapsto$ zfh2
TRANS	Transformations	Zn finger homeodomain 2 $\mapsto$ Zfh 2

Table 1: Gene text edit operations.

IR indexing and document weighting. Information retrieval (IR) provides a robust and efficient way to identify documents based on their content. We can conceive of lexical lookup as information retrieval for the purposes of this task by defining a document associated with a unique identifier to consist of the possible synonyms attributed to it. Jakarta Lucene[2] is an IR engine that implements $tf * idf$ scoring (i.e. each term in a query is given a weight according its term frequency–the number of times it occurs in a document–offset by its document frequency, which serves to devalue terms that are common in the overall corpus). The basic machinery is a reverse look-up table with the added value of term weighting to order possible documents matching a query. We created indexes from the synonym lists provided for the task. Documents were weighted by the frequency of their identifiers in the training corpus using add-one smoothing to assign non-zero weights to unseen identifiers.

[2] http://jakarta.apache.org/lucene/docs/index.html.

2.3. Identifier Disambiguation

Identifier disambiguation refers to the process of removing incorrect identifiers in cases where the lookup methods identify more than one possible identifier for a piece of text. While a clever system might attempt to identify cases where multiple tokens of the same string within a document refer to different entities, we simplified this sub-problem to the task of choosing a single identifier from a list of possibilities for a piece of text. We experimented with 3 approaches.

Modelling identifier co-occurrence. Gene identifier co-occurrence statistics were obtained from the training data. and the identifier with the maximum summed co-occurrence with the unambiguous identifiers obtained from the test document was selected. Ties were broken by backing off to maximum occurrence within the training corpus, and unresolved ties at this point were discarded.

IR query creation and term weighting. IR queries against the identifier documents can be used for disambiguation by creating a filter to limit the search to documents associated with pre-determined potential identifiers. Term weighting helps to distinguish identifiers retrieved by the IR engine based on the semantic importance of the individual tokens within the tagged text from the abstract. The weighting used (Table 2) is roughly based on the semantic classification of tokens outlined in (Hanisch et al., 2003).

Tok Class	PUNCTUATION	INDICI	INDICII
Examples	`-(),`	`gene, allele`	`family, locus`
Weight	0.5	1	10

Tok Class	SPECIFIER	COMMON	GENE
Examples	`2,II,`β	`she,from`	`Zfh,clk`
Weight	15	NF	20

Table 2: Breakdown of query term weighting. (NF (normalised frequency) for common words is computed by $10 - 9 \times \frac{TokenFreq}{HighestTokenFreq}$, giving a normalised weight between 1 and 10 where less frequent words are weighted higher.)

Heuristic approaches we identified several heuristics that help distinguish the correct identifier in some cases (Table 3). First, we observed that identifiers which have been unambiguously found elsewhere in the same document should be preferred (SD). It was also observed that identifiers in whose synonym list the tagged text occurs more often (ignoring case and expanding/collapsing acronyms) are better candidates (Reps). Finally, as a last resort that proved especially effective for yeast, we observed that the tagged text is more likely to occur earlier in the synonym list for the correct identifier (Ord).

3. Results

Several systems were submitted to the BioCreAtIvE evaluation. The systems have accuracy measures falling pretty much right on the median among all competitors. Section 3.1. details the configurations of the our submissions. Section 3.2. presents two baseline systems. And, Section 3.3. discusses some sources of error.

HEURISTIC	DESCRIPTION
SD	Prefer identifiers already found unambiguously in **s**ame **d**ocument.
Reps	Prefer identifiers with more **rep**eats of tagged text in synonym list.
Ord	Prefer identifiers for which the tagged text occurs earlier in synonym list **ord**er.

Table 3: Heuristic approaches to disambiguation.

3.1. Submissions

Table 4 describes the final configurations for the submitted systems. The first submission uses only information retrieval techniques for lookup and disambiguation. The system treats the gene text identified by the tagger as query input. Lucene is used to build and search an index of the synonym lists associated with unique gene identifiers.

The second submission uses the organism-specific tagger discussed in Section 2.1. for mouse and fly. For yeast a task 1A tagger is used incorporating only word and POS features. Different combinations of edit operations and disambiguation approaches are used for each organism. The third submission followed a unified approach for all three organisms, using organism-specific taggers, regular expression lookup, and co-occurrence disambiguation.

3.2. Baseline

We use a simple look-up table approach to produce two baselines for the grounding task. The first uses a tagger trained on the task 1A data with only word and POS features (Table 5) and the second uses a tagger trained on noisy data created from the synonym lists provided for the task (Table 6). The INC columns indicate whether all matches are included (+) or excluded (-) in cases where the tagged gene text matches into more than one synonym list.

ORG	INC	PRECISION	RECALL	$F_{\alpha=0.5}$
fly	+	18.4	30.0	22.8
	-	77.3	22.9	35.3
mouse	+	74.8	42.2	53.9
	-	84.8	37.2	51.7
yeast	+	92.9	35.2	51.1
	-	96.7	33.7	50.0

Table 5: A simple hash look-up with no edit operations using a tagger trained on task 1A data with word and POS features.

These scores reflect a couple of properties of the organism databases. First, polysemy and common words occur frequently among fly terminology relative to the other task 1B organisms. This is reflected in the dramatic decrease in precision when all ambiguous matches are included. Second, the yeast research community has the most well-defined and strictly followed naming conventions which is reflected in the relatively high baseline scores and the relatively small change when including or excluding all ambiguous matches.

SUBMISSION	GENE ID	LOOKUP	DISAMBIGUATION	DEVTEST $P/R/F$			EVAL $P/R/F$		
1. fly	Org-specific	IR	IR	60.6	67.7	64.0	59.2	74.8	66.1
mouse	Org-specific	IR	IR	92.2	72.0	61.0	77.0	59.6	67.2
yeast	Org-specific	IR	IR	68.0	55.3	80.9	94.8	72.1	81.9
2. fly	Org-specific	Edit:Cap	SD,IR,Ord	77.3	53.4	63.2	65.9	57.1	61.2
mouse	Org-specific	Edit:Cap,Ex,Wr,Punc,Trans	IR,Ord	75.3	65.6	70.1	81.3	67.3	73.6
yeast	W&POS 1A	Edit:Cap,Ex,Wr,Punc,Trans	SD,Reps,IR,Ord	89.5	81.6	85.4	91.5	79.0	84.8
3. fly	Org-specific	Regular expression	Co-occurrence	73.8	52.9	61.6	69.3	53.1	60.2
mouse	Org-specific	Regular expression	Co-occurrence	71.8	47.7	57.32	80.1	50.4	61.9
yeast	Org-specific	Regular expression	Co-occurrence	90.7	78.9	84.39	96.9	75.4	84.8

Table 4: Configuation and scores for final submissions. Best scores for development and evaluation testing (highlighted in bold) illustrates variation of best approaches for different organisms.

ORG	INC	PRECISION	RECALL	$F_{\alpha=0.5}$
fly	+	16.1	77.6	26.6
	-	78.0	44.4	56.6
mouse	+	71.6	50.3	59.1
	-	81.3	44.9	57.9
yeast	+	92.1	49.4	64.3
	-	94.7	47.9	63.6

Table 6: A simple look-up table approach with no edit operations using the organism-specific tagger with word, POS and gazetteer features.

3.3. Discussion

Types of errors. There are three main sources of error corresponding to the main sub-tasks (Section 1.), the first two primarily causing low recall problems and the third primarily affecting precision.

First, the tagger is not specific as to what it is identifying. Both the task 1A tagger and the organism-specific tagger are trained on gene names and some protein names. They are not trained to recognise all alleles or products, for instance, that may constitute an implicit gene mention. Nor do they recognise protein complexes that constitute an implicit mention of multiple genes.

Second, the synonym lists are not exhaustive and so we have cases where correctly tagged gene text does not match any synonyms. Even with the fuzzy matching approaches, we still miss some variations on gene terminology. (With respect to the organism-specific taggers, however, as the NER system should only identify text fragments contained within the synonym lists.)

Third, we have disambiguation errors where the tagged text matches more than one synonym list. This is illustrated in the dramatic difference between precision scores for fly, which has very loose naming conventions, and yeast, which has very strict naming conventions.

Estimating tagger performance. Returning all gene identifiers corresponding to *all* text fragments matching known synonyms shows an upper bound on recall from using only the provided synonym lists. Returning all gene ids for text fragments returned by the organism-specific taggers reveal a loss in recall of 6.7% for fly, 10.1% for mouse, and 0.8% for yeast from using more precise named-entity recognition (Table 7).

ORGANISM	REGEXP	TAGGER	LOSS
fly	90.1	83.4	6.7
mouse	78.9	68.8	10.1
yeast	81.6	80.8	0.8

Table 7: Recall of Organism-Specific Taggers

We also determined what percentage of the text fragments were valid synonyms (correspond to a gene ID in the gold standard answers) to get an estimation of the precision of the NER system (Table 8). This is the upper bound on the precision attainable by the disambiguation system at maximum recall. At the cost of recall, the precision of the disambiguation systems was often higher than this bound from the exclusion of matches that could not be disambiguated. The precision of the organism-specific taggers compares favorably with the baseline (all gene text fragments were returned with results filtered by POS tags), justifying the relatively small loss in recall.

ORGANISM	BASELINE	TAGGER	GAIN
fly	30.2	70.2	40.0
mouse	31.8	68.3	36.5
yeast	91.6	93.5	1.9

Table 8: Precision of Organism-Specific Taggers

Although there is room for improvement in the precision and recall of the NER component, there is a considerable loss in recall during the grounding process, indicating that the disambiguation methods are not optimal either. For the development set, the best performing systems on fly, mouse, and yeast (using the organism-specific taggers) had recalls of 67.7, 65.6, and 78.9 respectively, corresponding to losses of 15.7, 3.2, and 1.9 percent respectively.

4. Related work

There has been considerable interest in the automatic processing of biomedical documents recently due to the vast amounts of new information being published in the field. Recent workshops on the topic such as the Special Session on Language Processing and Biological Data at the 2002 Human Language Technology Workshop and the 2002 and 2003 Workshops on Natural Language Processing

in the Biomedical Domain[3] reflect this trend.

BioCreAtIvE is not the first evaluation event to be held in field of biomedical text mining. The 2002 KDD Cup at the International Conference on Knowledge Discovery and Data Mining included two tasks that involved data mining in molecular biology domains.[4] And the 2003 Text REtrieval Conference (TREC) included a Genomics Track[5] whose purpose is to study retrieval tasks involving genomics data. This task will also feature at TREC 2004.

Although similar, this work differs from the above mentioned tasks in several ways. Whereas the KDD Cup task provided a set of papers and a list of gene mentions, and asked the system to determine "whether the paper meets the Flybase gene-expression curation criteria, and for each gene, indicate whether the full paper has experimental evidence for gene products (RNA and/or protein)", this work is concerned with prerequisite step of identifying and normalizing those gene mentions. The information extraction component of the TREC 2003 task concerns automatically reproducing GeneRIF annotations providing topic summaries for MEDLINE records, and not the extraction of the gene mentions in the documents per se.

The issue of different textual realisations of gene terminology has been addressed in recent work. Yu and Agichtein (2003), for example, discuss methods for automatic extraction of gene and protein synonyms from text, a problem which involves a grounding task to determine which identified genes and proteins are synonyms of each other. Osborne et al. (2003) discuss methods for automatic generation of gene synonym variants as a means to expand queries for a document retrieval task. Hanisch et al. (2003) present a semi-automatic methodology for building curated gene dictionaries for named entity recognition.

There has also been recent work concerning grounding of biological named entities. Leidner et al. (2003) compare research in grounding spatial named entities (e.g with respect to world atlases and gazetteers) to grounding of biomedical spatial named entities (e.g with respect to brain or body atlases).

Most relevant is work by Hirschman et al. (2002), which discusses a problem very similar to BioCreAtIvE task 1B. The main thrust of the paper is gene entity recognition, however, as they are performing it in a domain without a gold standard, they treat the grounding task as a means to perform an extrinsic evaluation of named entity recognition. They present a grounding system similar to our baseline which illustrates the same precision/recall tradeoff for fly when including/excluding ambiguous matches.

Finally, Morgan et al. (2003) present a methodology for bootstrapping a gene named entity tagger using lists of curated genes for FlyBase to generate noisy training data that is similar to the process we used to create our organism-specific taggers.

[3]http://www.ccs.neu.edu/home/futrelle/bionlp/acl02/BIO/, http://www-tsujii.is.s.u-tokyo.ac.jp/ACL03/bionlp.htm

[4]http://www.biostat.wisc.edu/ craven/kddcup/

[5]http://trec.nist.gov/

5. Conclusion and Future Work

We have presented an evaluation of several approaches to grounding gene mentions with respect to gene database identifiers for fly, mouse, and yeast. The approaches address explicit gene mentions but fall short of fully addressing implicit mentions. The systems presented have accuracy measures falling pretty much right on the median among all BioCreAtIvE task 1B submissions.

Some initial error analysis has helped to tease apart tagging and grounding errors and differences among organism naming conventions. Precision and recall estimates were obtained for organism-specific gene named entity taggers bootstrapped from gene synonym lists and task 1B training materials. Ongoing analysis is looking into more specific grounding errors, including division of errors into the three categories mentioned in Section 3.3. A question that requires further exploration is why different tagging, lookup and disambiguation approaches worked better for different organisms, especially with respect to system 2 (Table 4).

Future system develeopment should account for gene mentions implicit in mentions of gene mutants, alleles, and products, which are a significant source of low recall in our current system.

6. Acknowledgments

This work was performed as part of the SEER project, which is supported by Scottish Enterprise Edinburgh-Stanford Link Grant R36759.

7. References

Jenny Finkel, Shipra Dingare, Christopher Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. 2004. Recognizing genes and proteins in MEDLINE abstracts. In *Proceedings of BioCreAtIvE 2004 (to appear)*, Granada, Spain, March.

Daniel Hanisch, Juliane Fluck, Heinz-Theodor Mevissen, and Ralf Zimmer. 2003. Playing biology's name game: Identifying protein names in scientific text. In *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB 2003)*, Lihue, Hawaii, USA, January.

Lynette Hirschman, Alexander A. Morgan, and Alexander S. Yeh. 2002. Rutabaga by any other name: extracting biological names. *Journal of Biomedical Informatics*, 35(4):247–259.

Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. 2003. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT/NAACL'03 Workshop on the Analysis of Geographic References*, Edmonton, Alberta, Canada, May.

Alex Morgan, Lynette Hirschman, Alexander Yeh, and Marc Colosimo. 2003. Gene name extraction using flybase resources. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, Sapporo, Japan, July.

Miles Osborne, Jeffrey Chang, Mark Cumiskey, Nipun Mehra, Veronica Rotemberg, Gail Sinclair, Matthew Smillie, Russ B. Altman, and Bonnie Webber. 2003. Edinburgh-stanford TREC 2003 genomics track: Notebook paper. In *Proceedings of the Twelfth Text REtrieval Conference*, Gaithersburg, Maryland, USA, November.

Hong Yu and Eugene Agichtein. 2003. Extracting synonymous gene and protein terms from biological literature. *Journal of Bioinformatics*, 19(1):i340–i349.