Bootstrapping Biomedical Ontologies for Scientific Text using NELL

Dana Movshovitz-Attias
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
dma@cs.cmu.edu

William W. Cohen
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
wcohen@cs.cmu.edu

Abstract

We describe an open information extraction system for biomedical text based on NELL (the Never-Ending Language Learner) (Carlson et al., 2010), a system designed for extraction from Web text. NELL uses a coupled semi-supervised bootstrapping approach to learn new facts from text, given an initial ontology and a small number of "seeds" for each ontology category. In contrast to previous applications of NELL, in our task the initial ontology and seeds are automatically derived from existing resources. We show that NELL's bootstrapping algorithm is susceptible to ambiguous seeds, which are frequent in the biomedical domain. Using NELL to extract facts from biomedical text quickly leads to semantic drift. To address this problem, we introduce a method for assessing seed quality, based on a larger corpus of data derived from the Web. In our method, seed quality is assessed at each iteration of the bootstrapping process. Experimental results show significant improvements over NELL's original bootstrapping algorithm on two types of tasks: learning terms from biomedical categories, and named-entity recognition for biomedical entities using a learned lexicon.

1 Introduction

NELL (the Never-Ending Language Learner) is a semi-supervised learning system, designed for extraction of information from the Web. The system uses a coupled semi-supervised bootstrapping approach to learn new facts from text, given an initial ontology and a small number of "seeds", *i.e.*, labeled examples for each ontology category. The new facts are stored in a growing structured knowledge base.

One of the concerns about gathering data from the Web is that it comes from various un-authoritative sources, and may not be reliable. This is especially true when gathering scientific information. In contrast to Web data, scientific text is potentially more reliable, as it is guided by the peer-review process. Open access scientific archives make this information available for all. In fact, the production rate of publicly available scientific data far exceeds the ability of researchers to "manually" process it, and there is a growing need for the automation of this process.

The biomedical field presents a great potential for text mining applications. An integral part of life science research involves production and publication of large collections of data by curators, and as part of collaborative community effort. Prominent examples include: publication of genomic sequence data, *e.g.*, by the Human Genome Project; online collections of three-dimensional coordinates of protein structures; and databases holding data on genes. An important resource, initiated as a means of enforcing data standardization, are ontologies describing biological, chemical and medical terms. These are heavily used by the research community. With this wealth of available data the biomedical field holds many information extraction opportunities.

We describe an open information extraction system adapting NELL to the biomedical domain. We present an implementation of our approach, named *BioNELL*, which uses three main sources of information: (1) a public corpus of biomedical scientific text, (2) commonly used biomedical ontologies, and

High PMI Seeds			Random Seeds		
SoxN	achaete	cycA	cac	section 33	28
Pax-6	Drosomycin	Zfh-1	crybaby	hv	Bob
BX-C	Ultrabithorax	GATAe	ael	LRS	dip
D-Fos	sine oculis	FMRFa	chm	sht	3520
Abd-A	dCtBP	Antp	M-2	AGI	tou
PKAc	huckebein	abd-A	shanti	disp	zen
Hmgcr	Goosecoid	knirps	Buffy	Gap	Scm
fkh	decapentaplegic	Sxl	lac	Mercurio	REPO
abdA	naked cuticle	BR-C	subcosta	mef	Ferritin
zfh-1	Kruppel	hmgcr	Slam	dad	dTCF
tkv	gypsy insulator	Dichaete	Cbs	Helicase	mago
CrebA	alpha-Adaptin	Abd-B	Sufu	ora	Pten
D-raf	doublesex	gusA	pelo	vu	sb
MtnA	FasII	AbdA	sombre	domain II	TrpRS
Dcr-2	GAGA factor	dTCF	TAS	CCK	ripcord
fushi tarazu	kanamycin resistance	Ecdysone receptor	GABAA receptor	diazepam binding inhibitor	yolk protein
Tkv	dCBP		Debcl	arm	

Table 1: Two samples of fruit-fly genes, taken from the complete fly gene dictionary. *High PMI Seeds* are the top 50 terms selected using PMI ranking, and *Random Seeds* are a random draw of 50 terms from the dictionary. These are used as seeds for the *Fly Gene* category (Section 4.2). Notice that the random set contains many terms that are often not used as genes including *arm*, *28*, and *dad*. Using these as seeds can lead to semantic drift. In contrast, high PMI seeds exhibit much less ambiguity.

(3) a corpus of Web documents.

NELL's ontology, including categories and seeds, has been manually designed during the system development. Ontology design involves assembling a set of interesting categories, organized in a meaningful hierarchical structure, and providing representative seeds for each category. Redesigning a new ontology for a technical domain is difficult without non-trivial knowledge of the domain. We describe a process of merging source ontologies into one structure of categories with seed examples.

However, as we will show, using NELL's bootstrapping algorithm to extract facts from a biomedical corpus is susceptible to noisy and ambiguous terms. Such ambiguities are common in biomedical terminology (see examples in Table 1), and some ambiguous terms are heavily used in the literature. For example, in the sentence "We have cloned an induced *white* mutation and characterized the insertion sequence responsible for the mutant phenotype", *white* is an ambiguous term referring to the name of a gene. In NELL, ambiguity is limited using coupled semi-supervised learning (Carlson et al., 2009): if two categories in the ontology are declared mutually exclusive, instances of one category are used as negative examples for the other, and the two categories cannot share any instances. To resolve the ambiguity of *white* with mutual exclusion, we would have to include a *Color* category in the ontology, and declare it mutually exclusive with the *Gene* category. Then, instances of *Color* will not be able to refer to genes in the KB. It is hard to estimate what additional categories should be added, and building a "complete" ontology tree is practically infeasible.

NELL also includes a polysemy resolution component that acknowledges that one term, for example *white*, may refer to two distinct concepts, say a color and a gene, that map to different ontology categories, such as *Color* and *Fly Gene* (Krishnamurthy and Mitchell, 2011). By including a *Color* category, this component can identify that *white* is both a color and a gene. The polysemy resolver performs word sense induction and synonym resolution based on relations defined between categories in the ontology, and labeled synonym examples. However, at present, BioNELL's ontology does not contain relation definitions (it is based only on categories), so we cannot include this component in our experiments. Additionally, it is unclear how to avoid the use of polysemous terms as category seeds, and no method has been suggested for selecting seeds that are representative of a single specific category.

To address the problem of ambiguity, we introduce a method for assessing the desirability of noun phrases to be used as seeds for a specific target category. We propose ranking seeds using a Pointwise Mutual Information (PMI) -based collocation measure for a seed and a category name. Collocation is measured based on a large corpus of domain-independent data derived from the Web, accounting for uses of the seed in many different contexts.

NELL's bootstrapping algorithm uses the morphological and semantic features of seeds to propose new facts, which are added to the KB and used as seeds in the next bootstrapping iteration to learn more facts. This means that ambiguous terms may be added at any learning iteration. Since *white* really *is* a name of a gene, it is sometimes used in the same semantic context as other genes, and may be added to the KB despite not being used as an initial seed.

To resolve this problem, we propose measuring seed quality in a *Rank-and-Learn* bootstrapping methodology: after every iteration, we rank all the instances that have been added to the KB by their quality as potential category seeds. Only high-ranking instances are used as seeds in the next iteration. Low-ranking instances are stored in the KB and "remembered" as true facts, but are not used for learning new information. This is in contrast to NELL's approach (and most other bootstrapping systems), in which there is no distinction between acquired facts, and facts that are used for learning.

2 Related Work

Biomedical Information Extraction systems have traditionally targeted recognition of few distinct biological entities, focusing mainly on genes (*e.g.*, (Chang et al., 2004)). Few systems have been developed for fact-extraction of many biomedical predicates, and these are relatively small scale (Wattarujeekrit et al., 2004), or they account for limited sub-domains (Dolbey et al., 2006). We suggest a more general approach, using bootstrapping to extend existing biomedical ontologies, including a wide range of sub-domains and many categories. The current implementation of BioNELL includes an ontology with over 100 categories. To the best of our knowledge, such large-scale biomedical bootstrapping has not been done before.

Bootstrap Learning and Semantic Drift. Carlson *et al.* (2010) use coupled semi-supervised bootstrap learning in NELL to learn a large set of category classifiers with high precision. One drawback of using iterative bootstrapping is the sensitivity of this method to the set of initial seeds (Pantel et al., 2009). An ambiguous set of seeds can lead to *semantic drift*, *i.e.*, accumulation of erroneous terms and contexts when learning a semantic class. Strict bootstrapping environments reduce this problem by adding boundaries or limiting the learning process, including learning mutual terms and contexts (Riloff and Jones, 1999) and using mutual exclusion and negative class examples (Curran et al., 2007).

McIntosh and Curran (2009) propose a metric for measuring the semantic drift introduced by a learned term, favoring terms different than the recent *m* learned terms and similar to the first *n*, (shown for n=20 and n=100), following the assumption that semantic drift develops in late bootstrapping iterations. As we will show, for biomedical categories, semantic drift in NELL occurs within a handful of iterations (< 5), however according to the authors, using low values for *n* produces inadequate results. In fact, selecting effective *n* and *m* parameters may not only be a function of the data being used, but also of the specific category, and it is unclear how to automatically tune them.

Seed Set Refinement. Vyas *et al.* (2009) suggest a method for reducing ambiguity in seeds provided by human experts, by selecting the tightest seed clusters based on context similarity. The method is described for an order of 10 seeds, however, in an ontology containing hundreds of seeds per class, it is unclear how to estimate the correct number of clusters to choose from. Another approach, suggested by Kozareva *et al.* (2010), is using only constrained contexts where both seed and class are present in a sentence. Extending this idea, we consider a more general collocation metric, looking at entire documents including both the seed and its category.

3 Implementation

3.1 NELL's Bootstrapping System

We have implemented BioNELL based on the system design of NELL. NELL's bootstrapping algorithm is initiated with an input ontology structure of categories and seeds. Three sub-components operate to introduce new facts based on the semantic and morphological attributes of known facts. At every iteration, each component proposes candidate facts, specifying the supporting evidence for each candidate, and the candidates with the most strongly supported evidence are added to the KB. The process and sub-components are described in detail by Carlson *et al.* (2010) and Wang and Cohen (2009).

3.2 Text Corpora

PubMed Corpus: We used a corpus of 200K full-text biomedical articles taken from the PubMed Central Open Access Subset (extracted in October 2010)[1], which were processed using the OpenNLP package[2]. This is the main BioNELL corpus and it

[1] http://www.ncbi.nlm.nih.gov/pmc/

[2] http://opennlp.sourceforge.net

is used to extract category instances in all the experiments presented in this paper.

Web Corpus: BioNELL's seed-quality collocation measure (Section 3.4) is based on a domain-independent Web corpus, the English portion of the ClueWeb09 data set (Callan and Hoy, 2009), which includes 500 million web documents.

3.3 Ontology

BioNELL's ontology is composed of six base ontologies, covering a wide range of biomedical subdomains: the Gene Ontology (GO) (Ashburner et al., 2000), describing gene attributes; the NCBI Taxonomy for model organisms (Sayers et al., 2009); Chemical Entities of Biological Interest (ChEBI) (Degtyarenko et al., 2008), a dictionary focused on small chemical compounds; the Sequence Ontology (Eilbeck et al., 2005), describing biological sequences; the Cell Type Ontology (Bard et al., 2005); and the Human Disease Ontology (Osborne et al., 2009). Each ontology provides a hierarchy of terms but does not distinguish concepts from instances.

We used an automatic process for merging base ontologies into one ontology tree. First, we group the ontologies under one hierarchical structure, producing a tree of over 1 million entities, including 856K terms and 154K synonyms. We then separate these into *potential categories* and *potential seeds*. *Categories* are nodes that are unambiguous (have a single parent in the ontology tree), with at least 100 descendants. These descendants are the category's *Potential seeds*. This results in 4188 category nodes. In the experiments of this paper we selected only the top (most general) 20 categories in the tree of each base ontology. We are left with 109 final categories, as some base ontologies had less than 20 categories under these restrictions. Leaf categories are given seeds from their descendants in the full tree of all terms and synonyms, giving a total of around 1 million potential seeds. Seed set refinement is described below. The seeds of leaf categories are later extended by the bootstrapping process.

3.4 BioNELL's Bootstrapping System

3.4.1 PMI Collocation with the Category Name

We define a seed quality metric based on a large corpus of Web data. Let s and c be a seed and a target category, respectively. For example, we can take s = "white", the name of a gene of the fruit-fly, and c = "fly gene". Now, let D be a document corpus (Section 3.2 describes the Web corpus used for ranking), and let D_c be a subset of the documents containing a mention of the category name. We measure the collocation of the seed and the category by the number of times s appears in D_c, $|Occur(s, D_c)|$. The overall occurrence of s in the corpus is given by $|Occur(s, D)|$. Following the formulation of Church and Hanks (1990), we compute the PMI-rank of s and c as

$$\text{PMI}(s, c) = \frac{|Occur(s, D_c)|}{|Occur(s, D)|} \qquad (1)$$

Since this measure is used to compare seeds of the same category, we omit the $\log$ from the original formulation. In our example, as *white* is a highly ambiguous gene name, we find that it appears in many documents that do not discuss the fruit fly, resulting in a PMI rank close to 0.

The proposed ranking is sensitive to the descriptive name given to categories. For a more robust ranking, we use a combination of rankings of the seed with several of its ancestors in the ontology hierarchy. In (Movshovitz-Attias and Cohen, 2012) we describe this hierarchical ranking in more detail and additionally explore the use of the binomial log-likelihood ratio test (BLRT) as an alternative collocation measure for ranking.

We further note that some specialized biomedical terms follow strict nomenclature rules making them easily identifiable as category specific. These terms may not be frequent in general Web context, leading to a low PMI rank under the proposed method. Given such a set of high confidence seeds from a reliable source, one can enforce their inclusion in the learning process, and specialized seeds can additionally be identified by high-confidence patterns, if such exist. However, the scope of this work involves selecting seeds from an ambiguous source, biomedical ontologies, thus we do not include an analysis for these specialized cases.

3.4.2 Rank-and-Learn Bootstrapping

We incorporate PMI ranking into BioNELL using a *Rank-and-Learn* bootstrapping methodology. After every iteration, we rank all the instances that have been added to the KB. Only high-ranking instances

Learning System	Bootstrapping Algorithm	Initial Seeds	Corpus
BioNELL	Rank-and-Learn with PMI	PMI top 50	PubMed
NELL	NELL's algorithm	Random 50	PubMed
BioNELL+Random	Rank-and-Learn with PMI	Random 50	PubMed

Table 2: Learning systems used in our evaluation, all using the PubMed biomedical corpus and the biomedical ontology described in Sections 3.2 and 3.3.

are added to the collection of seeds that are used in the next learning iteration. Instances with low PMI rank are stored in the KB and are not used for learning new information. We consider a high-ranking instance to be one with PMI rank higher than 0.25.

4 Experimental Evaluation

4.1 Experimental Settings

4.1.1 Configurations of the Algorithm

In our experiments, we ran BioNELL and NELL with the following system configurations, all using the biomedical corpus and the ontology described in Sections 3.2 and 3.3, and all running 50 iterations, in order to evaluate the long term effects of ranking. Section 4.2 includes a discussion on the learning rate of the tested systems which motivates the reason for evaluating performance at the 50th iteration.

To expand a category we used the following systems, also summarized in Table 2: (1) the *BioNELL* system, using Rank-and-Learn bootstrapping (Section 3.4.2) initialized with the top 50 seeds using PMI ranking, (2) the *NELL* system, using NELL's original bootstrapping algorithm (Section 3.1) initialized with 50 random seeds from the category's potential seeds (NELL does not provide a seed selection method), and (3) in order to distinguish the contribution of Rank-and-Learn bootstrapping over ranking the initial seeds, we tested a third system, *BioNELL+Random*, using Rank-and-Learn bootstrapping initialized with 50 random seeds.

4.1.2 Evaluation Methodology

Using BioNELL we can learn *lexicons*, collections of category terms accumulated after running the system. One evaluation approach is to select a set of learned instances and assess their correctness (Carlson et al., 2010). This is relatively easy for data extracted for general categories like City or Sports Team. For example, it is easy to evaluate the statement "London is a City". This task becomes more difficult when assessing domain-specific facts such as "Beryllium is an S-block molecular entity" (in fact, it is). We cannot, for example, use the help of Mechanical Turk for this task. A possible alternative evaluation approach is asking an expert. On top of being a costly and slow approach, the range of topics covered by BioNELL is large and a single expert is not likely be able to assess all of them.

We evaluated lexicons learned by BioNELL by comparing them to available resources. Lexicons of gene names for certain species are available, and the Freebase database (Google, 2011), an open repository holding data for millions of entities, includes some biomedical concepts. For most biomedical categories, however, complete lexicons are scarce.

4.1.3 Data Sets

We compared learned lexicons to category *dictionaries*, lists of concept terms taken from the following sources, which we consider as a Gold Standard.

We used three lexicons of biomedical categories taken from Freebase: Disease (9420 terms), Chemical Compound (9225 terms), and Drug (3896 terms).

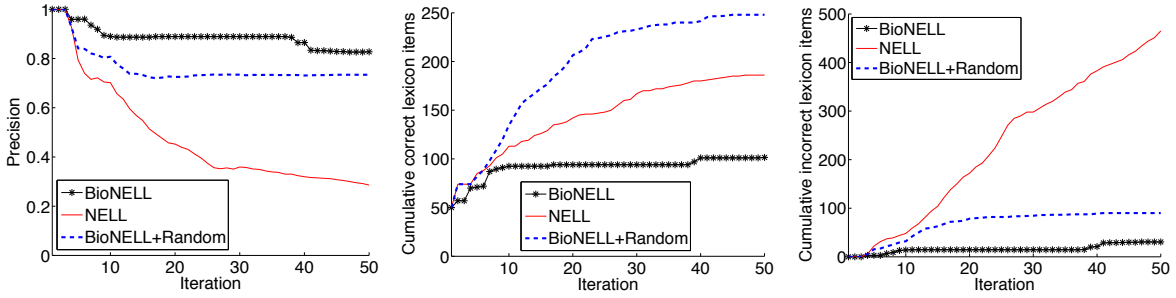
To evaluate gene names we used data from the BioCreative Challenge (Hirschman et al., 2005), an evaluation competition focused on annotations of genes and gene products. The data includes a dictionary of genes of the fruit-fly, *Drosophila Melanogaster*, which specifies a set of gene identifiers and possible alternative forms of the gene name, for a total of 7151 terms, which we consider to be the complete fly gene dictionary.

We used additional BioCreative data for a named-entity recognition task. This includes 108 scientific abstracts, manually annotated by BioCreative with gene IDs of fly genes discussed in the text. The abstracts contain either the gene ID or any gene name.

4.2 Extending Lexicons of Biomedical Categories

4.2.1 Recovering a Closed Category Lexicon

We used BioNELL to learn the lexicon of a closed category, representing genes of the fruit-fly,



(a) Precision (b) Cumulative correct items (c) Cumulative incorrect items

Figure 1: Performance per learning iteration for gene lexicons learned using BioNELL and NELL.

Learning System	Precision	Correct	Total
BioNELL	**.83**	109	132
NELL	.29	186	**651**
BioNELL+Random	.73	**248**	338
NELL $_{\text{by size } 132}$	.72	93	130

Table 3: Precision, total number of instances (*Total*), and correct instances (*Correct*) of gene lexicons learned with BioNELL and NELL. BioNELL significantly improves the precision of the learned lexicon compared with NELL. When examining only the first 132 learned items, BioNELL has both higher precision and more correct instances than NELL (last row, NELL $_{\text{by size } 132}$).

D. Melanogaster, a model organism used to study genetics and developmental biology. Two samples of genes from the full fly gene dictionary are shown in Table 1: *High PMI Seeds* are the top 50 dictionary terms selected using PMI ranking, and *Random Seeds* are a random draw of 50 terms. Notice that the random set contains many seeds that are not distinct gene names including *arm*, *28*, and *dad*. In contrast, high PMI seeds exhibit much less ambiguity. We learned gene lexicons using the test systems described in Section 4.1.1 with the high-PMI and random seed sets shown in Table 1. We measured the precision, total number of instances, and correct instances of the learned lexicons against the full dictionary of genes. Table 3 summarizes the results.

BioNELL, initialized with PMI-ranked seeds, significantly improved the precision of the learned lexicon over NELL (29% for *NELL* to 83% for *BioNELL*). In fact, the two learning systems using Rank-and-Learn bootstrapping resulted in higher precision lexicons (83%, 73%), suggesting that constrained bootstrapping using iterative seed ranking successfully eliminates noisy and ambiguous seeds.

BioNELL's bootstrapping methodology is highly restrictive and it affects the size of the learned lexicon as well as its precision. Notice, however, that while *NELL*'s final lexicon is 5 times larger than *BioNELL*'s, the number of correctly learned items in it are less than twice that of *BioNELL*. Additionally, *BioNELL+Random* has learned a smaller dictionary than *NELL* (338 and 651 terms, respectively) with a greater number of correct instances (248 and 186).

We examined the performance of *NELL* after the 7th iteration, when it has learned a lexicon of 130 items, similar in size to *BioNELL*'s final lexicon (Table 3, last row). After learning 130 items, *BioNELL* achieved both higher precision (83% versus 72%) and higher recall (109 versus 93 correct lexicon instances) than *NELL*, indicating that *BioNELL*'s learning method is overall more accurate.

After running for 50 iterations, all systems recover only a small portion of the complete gene dictionary (109-248 instances out of 7151), suggesting either that, (1) more learning iterations are required, (2) the biomedical corpus we use is too small and does not contain (frequent) mentions of some gene names from the dictionary, or (3) some other limitations exist that prevent the learning algorithm from finding additional class examples.

Lexicons learned using BioNELL show persistently high precision throughout the 50 iterations, even when initiated with random seeds (Figure 1A). By the final iteration, all systems stop accumulating further significant amounts of correct gene instances (Figure 1B). Systems that use PMI-based Rank-and-Learn bootstrapping also stop learning incorrect

Learning System	Precision			Correct			Total		
	CC	Drug	Disease	CC	Drug	Disease	CC	Drug	Disease
BioNELL	**.66**	**.52**	**.43**	63	508	276	96	972	624
NELL	.15	.40	.37	**74**	**522**	**288**	**449**	**1300**	**782**
NELL $_{\text{by size}}$	.58	.47	.37	58	455	232	100	968	623

Table 4: Precision, total number of instances (*Total*), and correct instances (*Correct*) of learned lexicons of *Chemical Compound* (CC), *Drug*, and *Disease*. BioNELL's lexicons have higher precision on all categories compared with NELL, while learning a similar number of correct instances. When restricting NELL to a total lexicon size similar to BioNELL's, BioNELL has both higher precision and more correct instances (last row, NELL $_{\text{by size}}$).

instances (*BioNELL* and *BioNELL+Random*; Figure 1C). This is in contrast to *NELL* which continues learning incorrect examples.

Interestingly, the highest number of correct gene instances was learned using Rank-and-Learn bootstrapping with random initial seeds (248 items; *BioNELL+Random*). This may happen when the random set includes genes that are infrequent in the general Web corpus, despite being otherwise category-specific in the biomedical context. As such, these would result in low PMI rank (see note in Section 3.4.1). However, random seed selection does not offer any guarantees on the quality of the seeds used, and therefore will result in unstable performance. Note that *BioNELL+Random* was initiated with the same random seeds as *NELL*, but due to the more constrained Rank-and-Learn bootstrapping it achieves both higher recall (248 versus 186 correct instances) and precision (73% versus 29%).

4.2.2 Extending Lexicons of Open Categories

We evaluated learned lexicons for three open categories, *Chemical Compound* (*CC*), *Drug*, and *Disease*, using dictionaries from Freebase. Since these are open categories — new drugs are being developed every year, new diseases are discovered, and varied chemical compounds can be created — the Freebase dictionaries are not likely to contain complete information on these categories. For our evaluation, however, we considered them to be complete.

We used *BioNELL* and *NELL* to learn these categories, and for all of them *BioNELL*'s lexicons achieved higher precision than *NELL* (Table 4). The number of correct learned instances was similar in both systems (63 and 74 for *CC*, 508 and 522 for *Drug*, and 276 and 288 for *Disease*), however in *BioNELL*, the additional bootstrapping restrictions assist in rejecting incorrect instances, resulting in a smaller, more accurate lexicon.

We examined *NELL*'s lexicons when they reached a size similar to *BioNELL*'s final lexicons (at the 8th, 42nd and 39th iterations for *CC*, *Drug*, and *Disease*, respectively). *BioNELL*'s lexicons have both higher precision and higher recall (more correct learned instances) than the comparable *NELL* lexicons (Table 4, NELL $_{\text{by size}}$, last row).

4.3 Named-Entity Recognition using a Learned Lexicon

We examined the use of gene lexicons learned with BioNELL and NELL for the task of recognizing concepts in free text, using a simple strategy of matching words in the text with terms from the lexicon. We use data from the BioCreative challenge (Section 4.1.3), which includes text abstracts and the IDs of genes that appear in each abstract. We show that BioNELL's lexicon achieves both higher precision and recall in this task than NELL's.

We implemented an *annotator* for predicting what genes are discussed in text, which uses a gene lexicon as input. Given sample text, if any of the terms in the lexicon appear in the text, the corresponding gene is predicted to be discussed in the text. Following BioCreative's annotation format, the annotator emits as output the set of gene IDs of the genes predicted for the sample text.

We evaluated annotators that were given as input: the complete fly-genes dictionary, a filtered version of that dictionary, or lexicons learned using BioNELL and NELL. Using these annotators we predicted gene mentions for all text abstracts in the data. We report the average precision (over 108 text

Lexicon	Precision	Correct	Total
BioNELL	.90	18	20
NELL	.02	5	268
BioNELL+Random	.03	3	82
Complete Dictionary	.09	153	1616
Filtered Dictionary	.18	138	675

Table 5: Precision, total number of predicted genes (*Total*), and correct predictions (*Correct*), in a named-entity recognition task using a complete lexicon, a filtered lexicon, and lexicons learned with BioNELL and NELL. BioNELL's lexicon achieves the highest precision, and makes more correct predictions than NELL.

abstracts) and number of total and correct predictions of gene mentions, compared with the labeled annotations for each text (Table 5).

Many gene names are shared among multiple variants. For example, the name *Antennapedia* may refer to several gene variations, *e.g.*, *Dgua\Antp* or *Dmed\Antp*. Thus, in our precision measurements, we consider a prediction of a gene ID as "true" if it is labeled as such by BioCreative, or if it shares a synonym name with another true labeled gene ID.

First, we used the complete fly gene dictionary for the recognition task. Any dictionary gene that is mentioned in the text was recovered, resulting in high recall. However, the full dictionary contains ambiguous gene names that contribute many false predictions to the complete dictionary annotator, leading to a low precision of 9%.

Some ambiguous terms can be detected using simple rules, *e.g.*, short abbreviations and numbers. For example, *section 9* is a gene named after the functional unit to which it belongs, and abbreviated by the symbol *9*. Clearly, removing *9* from the full lexicon should improve precision without great cost to recall. We similarly filtered the full dictionary, removing one- and two-letter abbreviations and terms composed only of non-alphabetical characters, leaving 6253 terms. Using the filtered dictionary, precision has doubled (18%) with minor compromise to recall. Using complete or manually refined gene dictionaries for named-entity recognition has been shown before to produce similar high-recall and low-precision results (Bunescu et al., 2005).

We evaluated annotators on gene lexicons learned with BioNELL and NELL. *BioNELL*'s lexicon achieved significantly higher precision (90%) than other lexicons (2%-18%). It is evident that this lexicon contains few ambiguous terms as it leads to only 2 false predictions. Note also, that *BioNELL*'s lexicon has both higher precision and recall than *NELL*.

5 Conclusions

We have proposed a methodology for an open information extraction system for biomedical scientific text, using an automatically derived ontology of categories and seeds. Our implementation is based on constrained bootstrapping in which seeds are ranked at every iteration.

The benefits of iterative seed ranking have been demonstrated, showing that our method leads to significantly less ambiguous lexicons for all the evaluated biomedical concepts. BioNELL shows 51% increase over NELL in the precision of a learned lexicon of chemical compounds, and 45% increase for a category of gene names. Importantly, when BioNELL and NELL learn lexicons of similar size, BioNELL's lexicons have both higher precision and recall. We have demonstrated the use of BioNELL's learned gene lexicon as a high precision annotator in an entity recognition task (with 90% precision). The results are promising, though it is currently difficult to provide a similar quantitative evaluation for a wider range of concepts.

Many interesting improvements could be made in the current system, mainly discovery of relations between existing ontology categories. In addition, we believe that Rank-and-Learn bootstrapping and iterative seed ranking can be beneficial in general, domain-independent settings, and we would like to explore further use of this method.

Acknowledgments

This work was funded by grant 1R101GM081293 from NIH, IIS-0811562 from NSF and by a gift from Google. The opinions expressed in this paper are solely those of the authors.

References

M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

J. Bard, S.Y. Rhee, and M. Ashburner. 2005. An ontology for cell types. *Genome Biology*, 6(2):R21.

R. Bunescu, R. Ge, R.J. Kate, E.M. Marcotte, and R.J. Mooney. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2).

J. Callan and M. Hoy. 2009. Clueweb09 data set. *http://boston.lti.cs.cmu.edu/Data/clueweb09/*.

A. Carlson, J. Betteridge, E.R. Hruschka Jr, T.M. Mitchell, and SP Sao Carlos. 2009. Coupling semi-supervised learning of categories and relations. *Semi-supervised Learning for Natural Language Processing*, page 1.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr, and T.M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, volume 2, pages 3–3.

J.T. Chang, H. Schütze, and R.B. Altman. 2004. Gapscore: finding gene and protein names one word at a time. *Bioinformatics*, 20(2):216.

K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

J.R. Curran, T. Murphy, and B. Scholz. 2007. Minimising semantic drift with mutual exclusion bootstrapping. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*.

K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. 2008. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl 1):D344.

A. Dolbey, M. Ellsworth, and J. Scheffczyk. 2006. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In *Proceedings of KR-MED*, pages 87–94. Citeseer.

K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner. 2005. The sequence ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):R44.

Google. 2011. Freebase data dumps. http://download.freebase.com/datadumps/.

L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. 2005. Overview of biocreative: critical assessment of information extraction for biology. *BMC bioinformatics*, 6(Suppl 1):S1.

Z. Kozareva and E. Hovy. 2010. Not all seeds are equal: measuring the quality of text mining seeds. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 618–626. Association for Computational Linguistics.

J. Krishnamurthy and T.M. Mitchell. 2011. Which noun phrases denote which concepts? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

T. McIntosh and J.R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 396–404. Association for Computational Linguistics.

D. Movshovitz-Attias and W.W. Cohen. 2012. Bootstrapping biomedical ontologies for scientific text using nell. Technical report, Carnegie Mellon University, CMU-ML-12-101.

J. Osborne, J. Flatow, M. Holko, S. Lin, W. Kibbe, L. Zhu, M. Danila, G. Feng, and R. Chisholm. 2009. Annotating the human genome with disease ontology. *BMC genomics*, 10(Suppl 1):S6.

P. Pantel, E. Crestan, A. Borkovsky, A.M. Popescu, and V. Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 938–947. Association for Computational Linguistics.

E. Riloff and R. Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-99)*, pages 474–479.

E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 37:5–15, Jan.

V. Vyas, P. Pantel, and E. Crestan. 2009. Helping editors choose better seed sets for entity set expansion. In *Proceeding of the 18th ACM conference on Information and knowledge management*. ACM.

R.C. Wang and W.W. Cohen. 2009. Character-level analysis of semi-structured documents for set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1503–1512. Association for Computational Linguistics.

T. Wattarujeekrit, P. Shah, and N. Collier. 2004. Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC bioinformatics*, 5(1):155.