Bootstrapping Biomedical Ontologies for Scientific Text using NELL

Dana Movshovitz-Attias William W. Cohen

May 2012
CMU-ML-12-101

Carnegie Mellon

Bootstrapping Biomedical Ontologies for Scientific Text using NELL

Dana Movshovitz-Attias **William W. Cohen**

May 2012
CMU-ML-12-101

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

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.

Keywords: Bootstrap Learning, Semi-Supervised Learning, Information Extraction, Pointwise Mutual Information

Abstract

We describe an open information extraction system for biomedical text based on NELL (the Never-Ending Language Learner) [7], 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) [7] 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 using data gathered 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. Data that comes from non-experts may be inaccurate. Sources of facts are not always cited and it is difficult to verify their integrity. The problem is amplified when a wrong fact, stated by one source, is repeated by others, like a "rumor". Detecting this type of duplicated information is not trivial, especially when the content is presented in varied forms.

In contrast to Web data, scientific text is potentially more reliable, as it is guided by the peer-review process. Facts in published papers are written by experts in their field. Not only that, claims are supported by experimental evaluations so that authors may convince their peers of the validity of their findings. Open access scientific archives make this information available for all, and they are continually updated with newly published materials. Other sources of public scientific data include databases of experimental results as well as human-curated structured information. In fact, the production rate of publicly available scientific data far exceeds the ability of researchers to "manually" process it, when they are searching for information. There is a growing need for automation of this process in a way that combines available resources.

The biomedical field presents a great potential for text mining applications. An integral part of Life Science research involves the production and publication of large collections of data by curators, and as part of a collaborative community effort. Prominent examples include: publication of genomic sequence data, for example, by the Human Genome Project; online collections of the three-dimensional coordinates of protein structures; and databases holding data on genes, including descriptions of gene functions, and the pathways in which they are involved (if known). These are updated by a wide community of researchers. An important biomedical resource, initiated as a means of enforcing data standardization, is the varied collection of ontologies describing biological, chemical and medical terms. These are maintained as part of large scale projects, spanning many years and considerable human effort, and are therefore heavily used by the research community. With this wealth of data available through online tools, databases, ontologies, and literature, the biomedical field holds many information extraction opportunities.

We describe an open information extraction system adapting NELL to the biomedical domain, using scientific resources available from the Web. 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) existing, commonly used biomedical ontologies, and (3) a corpus of Web documents.

NELL's ontology, including both categories and seeds, has been manually designed during the system development. Ontology design involves assembling a set of interesting categories, gathering these categories into a meaningful hierarchical structure, and providing representative examples (seeds) for each category. Redesigning a new ontology for a technical domain is difficult

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 genes of the fruit-fly, taken from the complete dictionary of fly genes. *High PMI Seeds* are the top 50 terms selected using PMI ranking, and *Random Seeds* are a random draw of 50 terms from the gene dictionary. These sets of genes are used as seeds for the *Fly Gene* category (described in Section 4.3). Notice that the random set contains many terms that are often not used as gene names including *arm*, *28*, and *dad*. Using these as seeds can lead to semantic drift. In contrast, high PMI seeds exhibit much less ambiguity.

without non-trivial knowledge of the domain. We describe an automatic process of merging source ontologies into one hierarchical structure of categories, with seed examples for every category. The ontologies we use cover a wide range of terms from biology, chemistry, and medicine, and they potentially allow for an interesting knowledge base to be acquired.

However, as we will show, using NELL's existing bootstrapping algorithm to extract facts from a biomedical corpus is highly susceptible to noisy and ambiguous terms. Such ambiguities are common in biomedical terminology (some examples can be seen in Table 1 and Figure 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 refers to the name of a gene, or more specifically, a gene mutation causing a white-eye phenotype in male flies. Using *white* in the KB, as an example of a gene, may eventually lead to learning that *green* and *gray* are also genes, and they may not be. In NELL, ambiguity is limited

A

Gene ID	Name 1	Name 2	Name 3
FBgn0000011	white	enhancer of garnet	e(g)
FBgn0002545	section 9	9	lf
FBgn0003204	raspberry	IMP dehydrogenase	ras-l
FBgn0004034	yellow	y	T6
FBgn0012326	Antp	Antennapedia	Dgua\Antp
FBgn0020493	dad	Daughters against dpp	Dad1

B

Abstract	Gene IDs
In Drosophila, MR (male recombination) second chromosomes are known to act as mutators and recombination inducers in males. The induction of visible mutations by MR is observed at only a limited number of genes, such as singed bristle (sn), raspberry eye colour (ras), yellow body colour (y) and a carmine eye colour (car) …	FBgn0003204 FBgn0004034

Figure 1: A sample from the BioCreative data set: (A) a list of gene identifiers (first column) as well as alternative common names and symbols used to describe each gene in the literature (second to last columns). The full data contains 7151 terms; and (B) sample abstract and two IDs of genes that have been annotated as being discussed in the text. In this example, the gene IDs *FBgn0003204* and *FBgn0004034* (can be found in the table) refer to the *raspberry* and *yellow* genes which are mentioned in the abstract. The full data contains 108 abstracts.

using coupled semi-supervised learning [6]: if two categories in the ontology are declared as mutually exclusive, instances of one category are used as negative examples for the other. This means that two mutually exclusive categories cannot share any instances. Thus, to resolve the ambiguity of the *white* gene using mutual exclusion, we would have to include a *Color* category somewhere in the ontology, and declare it mutually exclusive with the *Gene* category. Then, instances of *Color*, like *white* or *green*, will no longer 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 method for detecting and compensating for ambiguity. A polysemy resolution component has been added 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* [23]. By adding a *Color* category to the ontology, this component can identify that *white* is indeed polysemous. It is both a color and a gene. While polysemy resolution is an important ambiguity resolver in NELL, the question remains, what other overlapping categories could there be for names of genes, diseases or molecules? 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 knowledge base, and used as seeds in the next bootstrapping iteration to learn more facts. This means that ambiguous terms may be introduced into the system at any learning iteration. *White* really *is* a name of a gene, and it may very well be used in the same context as other genes that have more "traditional" names (such as, Helicase, SoxN or dTCF). An extraction system that is based on semantic context would be right in suggesting that *white* be added as a gene in the knowledge base, although it is more frequently used to name a color. To resolve this problem, we propose using seed quality measures in a *Rank-and-Learn* bootstrapping methodology. After every bootstrapping iteration, we rank all the instances that have been added to the knowledge base by their quality as potential category seeds. Only high-ranking instances are added to the collection of seeds that are used in the next bootstrapping iteration. Low-ranking instances are stored in the knowledge base and "remembered" as true facts, but they 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.

The rest of this paper is organized as follows. In Section 2 we review related work, including a review of the reasons for the high rate of ambiguity in biomedical terminology. Next, in Section 3, we present our implementation of BioNELL. We describe the data and ontologies that have been used, and we present our proposed seed quality collocation measure. An experimental evaluation of the system is given in Section 4, including demonstrated use-cases. We conclude that using ranking during bootstrapping significantly reduces ambiguity when learning biomedical concepts (Section 5).

2 Related Work

Biomedical Information Extraction systems have traditionally targeted recognition of few distinct biological entities [30], focusing mainly on genes and proteins [24, 10, 29, 9]. Few systems have been developed for fact-extraction of a larger set of biomedical predicates, and these are relatively small scale [34], or they account for limited biomedical sub-domains [16] or corpora concerning specific species [31]. 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.

Sources of Ambiguity in Biomedical Terminology. It has been shown that biomedical terminology suffers from a higher level of ambiguity than what is found in ordinary English words, with even greater ambiguity found in gene names [11, 22] (see examples in Table 1 and Figure 1). This problem is manifested in two main forms. The first is the use of short-form names, lacking meaningful morphological structure, including abbreviations of three or less letters as well as isolated numbers. The second is ambiguous and polysemous terms used to describe names of genes, organisms, and biological systems and processes. For examples, *peanut* is used as both the name of a plant and a gene, and many gene names are often shared across species. What's more, with

a limited possible number of three-English-letter abbreviations, and an estimate of around 35,000 human genes alone, newly introduced abbreviations are bound to overlap existing ones. Krallinger *et al.* [22] provide an in-depth review discussing the ambiguous nature of this domain-specific terminology in greater detail.

Bootstrap Learning and Semantic Drift. Carlson *et al.* [7] use a coupled semi-supervised bootstrap learning approach 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 [26]. An ambiguous set of seeds can lead to the problem of "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 [27] and using mutual exclusion and negative class examples [14]. In BioNELL, the initial seeds given to the bootstrapping system are taken from biomedical ontology terms that exhibit this high ambiguity. By refining the automatically derived set of initial seeds, we can remove ambiguous terms and minimize semantic drift.

Seed Set Refinement. Vyas *et al.* [32] suggest a method for reducing ambiguity in seeds provided by human experts, by selecting the K tightest clusters based on context similarity, for a pre-selected K. The method is described for groups in the order of 10 seeds. In a large ontology containing hundreds of potential seeds per class, it is unclear how to estimate the correct number of clusters to choose from. Another interesting approach, suggested by Kozareva *et al.* [21], is using only constrained contexts where both seed and class are present in the sentence. Extending this idea, we consider a more general collocation metric, looking at entire documents including both the seed and its category. According to this metric we rank the initial set of seeds and all learned facts, and we use the rank as a measure for their suitability to be used as seeds in later bootstrapping rounds.

Word Collocation. Various collocation measures are used in the context of information extraction, including pointwise mutual information (PMI) [13], the t-test [12], and binomial log-likelihood ratio test (BLRT) [17]. A review of the benefits and short-comings of several collocation methods can be found in [1]. We elaborate on the limitations of using BLRT for seed refinement in Section 3.4.3.

3 Implementation

We have implemented BioNELL based on the system design and bootstrapping approach of NELL. In this section we include a description of NELL's bootstrapping algorithm. We then describe the data used to build BioNELL, and describe a process for merging source ontologies into one ontology including seeds. Finally, we define our seed ranking metric, and present how it is used in BioNELL's bootstrapping process. We also describe an alternative collocation measure, which we compare with PMI.

3.1 NELL's Bootstrapping System

NELL's bootstrapping algorithm is initiated with an input ontology structure and *seeds*, labeled examples for every ontology category. These are used to populate a knowledge base of learned facts. Three underlying sub-components operate to suggest candidate facts to the knowledge base: One component extracts free text from the corpus using semantic patterns [8]; The second builds Web queries using currently known facts from the knowledge base, and mines the results for new candidate facts [33]; The final component classifies noun phrases according to their morphological attributes. At every iteration, each component proposes new candidate facts, specifying the supporting evidence for each candidate. Finally, the proposed candidates with the most strongly supported evidence are promoted and added to the knowledge base. With this process, the KB of facts grows. This process and all system sub-components are described in greater detail by Carlson *et al.* [7] and Wang and Cohen [33].

At present, the Web version of NELL has accumulated a knowledge base of 986K asserted instances of 266 categories and 199 relations.

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], processed using the OpenNLP package[2]. This is the main BioNELL corpus and it is used to extract category instances in all the experiments presented in this paper.

Web Corpus: BioNELL's seed-quality collocation measure (see Section 3.4) is based on a domain-independent Web corpus, the English portion of the ClueWeb09 data set [5], which includes 500 million web documents.

3.3 Ontology

BioNELL's ontology is composed of six base ontologies, covering a wide range of terms from biology, chemistry, and medicine: the Gene Ontology (GO) [2], describing gene attributes; the NCBI Taxonomy for model organisms [28]; Chemical Entities of Biological Interest (ChEBI) [15], a dictionary of molecular entities and small chemical compounds; the Sequence Ontology [18], describing biological sequences; the Cell Type Ontology [3]; and the Human Disease Ontology [25]. 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, as follows. First, we group the six ontologies under one hierarchical structure, producing a tree of over 1 million entities, including 856K terms and an additional 154K synonyms. We then separate these into *potential categories* and *potential seeds* for the ontology categories. *Categories* are nodes that are that are unambiguous (have a single parent in the ontology tree), and have at least 100 descendants — these descendants are the category's *Potential seeds*. This results in 4188 potential category

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

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

nodes. In the experiments of this paper we selected only the top (most general) 20 potential categories in the tree of each base ontology. We are left with 109 final categories, as some base ontologies had less than 20 potential 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.

The ontologies we have chosen are mutually exclusive with respect to the domains they cover. For this reason, categories from each base ontology are declared as mutually exclusive with the categories of every other base ontology. Within each base ontology, categories are mostly not mutually exclusive, with the exception of the top three categories of GO: Biological Process, Cellular Component, and Molecular Function. These three categories are treated as base ontologies for the purpose of mutual exclusion.

3.4 Extending BioNELL with Rank-and-Learn Bootstrapping

For each category in the BioNELL ontology we have at least a hundred potential seeds, derived from a base ontology definition, and many of them are used ambiguously in the biomedical literature. Using them as initial examples to ontology categories, and using NELL's bootstrapping algorithm to expand that ontology, results in a fast growing set of facts that are irrelevant to the category being learned (as is demonstrated in our evaluations below). We wish to define a method for assessing seed quality, based on a large corpus of data derived from the Web. Seeds are ranked according to their "quality", and this ranking is used in a *Rank-and-Learn* bootstrapping process, where only high-ranking seeds are incorporated in any further learning iterations. Below we use the term *seeds*, not only with reference to initial labeled examples for a category, but also to learned category instances that are used for learning and expanding a category at any of the subsequent bootstrapping steps.

3.4.1 PMI Collocation with the Category Name

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 [13], we compute the PMI-rank of s and c as

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

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. This intuitive and simple-to-calculate measure captures an important relationship between the category and seed, and our experiments show that using it alleviates many ambiguities.

We extend the seed rank definition by measuring the collocation of seeds with their three nearest ancestors in BioNELL's ontology tree. In other words, a *Fly Gene* is also a *Gene*, and this fact is captured in the ontology structure by the fact that the *Fly Gene* category is a descendant of the *Gene* category. We combine these ranks, placing an emphasis on collocation with the immediate ancestor, the category, by

$$\begin{aligned} &\text{combined-PMI}(s, c) = \\ &\quad \lambda_1 \cdot \text{PMI}(s, c) + \\ &\quad \lambda_2 \cdot \text{PMI}(s, A(c)) + \\ &\quad \lambda_3 \cdot \text{PMI}(s, A(A(c))) \end{aligned} \tag{2}$$

where $A(x)$ denotes the ancestor of x in the ontology structure, $\lambda_1 = \frac{1}{2}$, and $\lambda_2, \lambda_3 = \frac{1}{4}$. For categories with only a single ancestor the PMI ranks are averaged (effectively, $\lambda_2 = \frac{1}{2}$ and the third term is not used), and in the case of a category with no ancestors, only PMI(s, c) is used. BioNELL's ontology tree does not contain ambiguous categories with two parents (see Section 3.3). In the following evaluations we use the combined-PMI rank for seeds and categories.

3.4.2 Rank-and-Learn Bootstrapping

We incorporate PMI ranking into BioNELL using a *Rank-and-Learn* bootstrapping methodology. After every bootstrapping iteration, we rank all the new category instances that have been added to the knowledge base. Only high-ranking instances are added to the collection of seeds that are used in the next learning iteration. Instances with low PMI rank are stored in the knowledge base and "remembered" as true facts, but they are not used for learning any new information. Using this methodology, the bootstrapping system is initialized with an unambiguous set of category examples, and no further ambiguous examples are added to it at any point. The learning subcomponents of the system can then use a "clean" set of examples from which they infer meaningful morphological patterns and semantic context representative of the category. We consider a high-ranking instance to be one with PMI rank higher than 0.25, which means it has a high collocation rank with at least one of its early ancestors, or moderate collocation with the category itself.

3.4.3 Alternative Ranking Models Based on Binomial Log-Likelihood Ratio Test (BLRT)

We used the binomial log-likelihood ratio test (BLRT) [17] as an alternative collocation measure. We use it to compare the occurrence of a seed, s in two sets of documents, D_c and D (as defined above). The idea behind BLRT is to compare the ratio of occurrence of a word in two text corpora, while assuming an underlying binomial distribution of words. Two possible hypotheses are considered: (1) the two ratios are drawn from different distributions, and (2) from the same distribution.

The BLRT rank for a seed s is given by

$$\text{BLRT}(s, c) = 2 \log \frac{L(p_1, k_1, n_1) L(p_2, k_2, n_2)}{L(p, k_1, n_1) L(p, k_2, n_2)} \tag{3}$$

where

$$k_1 = |Occur(s, D_c)| \quad (4)$$
$$k_2 = |Occur(s, D)| \quad (5)$$
$$n_1 = |D_c| \quad (6)$$
$$n_2 = |D| \quad (7)$$
$$p_i = \frac{k_i}{n_i} \quad (8)$$
$$p = \frac{k_1 + k_2}{n_1 + n_2} \quad (9)$$
$$L(p, k, n) = p^k (1 - p)^{n-k} \quad (10)$$

The main drawback of using this approach is the symmetry in considering the two random variables being tested. Seeds that are highly frequent in the general corpus but not in the category corpus (*i.e.*, with $p_2 \gg p_1$) get a high score, simply because the ratios are very different. In viewing this rank as a measure of relevance of a seed to a category, we can assume that such seeds would make undesirable bootstrapping examples. To address this, we also consider a *modified-BLRT* rank where a seed with higher occurrence ratio in the general corpus ($p_2 > p_1$) gets rank 0.

4 Experimental Evaluation

We start this section with suggestions of possible use-cases of BioNELL as a knowledge source for two types of information extraction tasks: (1) extending a lexicon for a biomedical category, and (2) named-entity recognition for biomedical entities using a learned lexicon. These tasks are described in order to motivate our evaluation of the system. Next, we describe the experimental settings and evaluation process. Finally, we evaluate the system's performance over the two described tasks. Through these evaluations we give a qualitative measure of the benefits of using PMI seed ranking and Rank-and-Learn bootstrapping.

4.1 Use-Cases for BioNELL

BioNELL was designed to populate a KB of biomedical categories with facts extracted from scientific text. The process begins with a partial lexicon (the seeds) for each pre-defined concept (the categories). With every iteration, the lexicon of each concept is extended as new facts are added by the bootstrapping algorithm. At the end of every iteration, BioNELL contains a collection of lexicons of biomedical concepts organized in a hierarchical structure. These lexicons can be used for a variety of applications including search and data discovery tools.

As an example, a lexicon for a concept can be used to recognize this concept in free text. One simple strategy is matching words in the text with terms from the lexicon. Lexicons learned using BioNELL can be used for this task when no complete lexicons are available for a concept. In our evaluation we show that a gene lexicon learned with BioNELL is less ambiguous than a complete gene lexicon and therefore achieves higher precision at this recognition task.

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
BioNELL+BLRT	Rank-and-Learn with BLRT	BLRT top 50	PubMed
BioNELL+mBLRT	Rank-and-Learn with mBLRT	mBLRT top 50	PubMed

Table 2: Learning systems used in our evaluation, including the main system *BioNELL*, the original *NELL* system, and three additional baseline configurations. All of the tested systems use the PubMed biomedical corpus and the biomedical ontology described in Sections 3.2 and 3.3.

4.2 Experimental Settings

4.2.1 Configurations of the Algorithm

In our experiments, we ran BioNELL and NELL using the following system configurations (described below and summarized in Table 2), all using the biomedical corpus and the ontology described in Sections 3.2 and 3.3. All systems ran for 50 iterations, in order to evaluate the long term effects of ranking on the KB. Section 4.3 includes a discussion on the learning rate of the tested systems which motivates the reason for evaluating performance at the 50th iteration.

Under each system configuration we distinguish a test category for which we assess the quality of the instances predicted by the system, comparing it against a Gold Standard dictionary. The set of seeds used to initialize the test category as well as the bootstrapping algorithm used for expansion are described below. The rest of the categories are initialized with a random set of seeds and expanded with the baseline bootstrapping algorithm of NELL. This testing methodology allows to evaluate the effect of ranking on one category in isolation of the rest of the ontology.

To expand the test category we used the following main systems: (1) the *BioNELL* system, which uses Rank-and-Learn bootstrapping (see Section 3.4.2) initialized with the top 50 seeds using PMI ranking with the category name, and (2) the *NELL* system, which uses NELL's original bootstrapping algorithm (see Section 3.1 and [7] for more details) initialized with a random set of 50 seeds from the category's potential seeds (NELL does not provide a seed selection method). In order to distinguish the contribution of Rank-and-Learn bootstrapping over ranking the initial seeds, we tested a third system, *BioNELL+Random*, using BioNELL's Rank-and-Learn bootstrapping initialized with 50 random seeds. As an alternative to the PMI ranking model, we tested two additional systems using BioNELL's bootstrapping methodology where PMI ranks were replaced with BLRT and modified-BLRT ranks (see Section 3.4.3). These are named *BioNELL+BLRT* and *BioNELL+mBLRT*. Table 2 contains a succinct summary of all configurations.

4.2.2 Evaluation Methodology

Using BioNELL we can learn *lexicons*, collections of terms, for categories in the ontology. A *lexicon* is a collection of category instances learned after using the system.

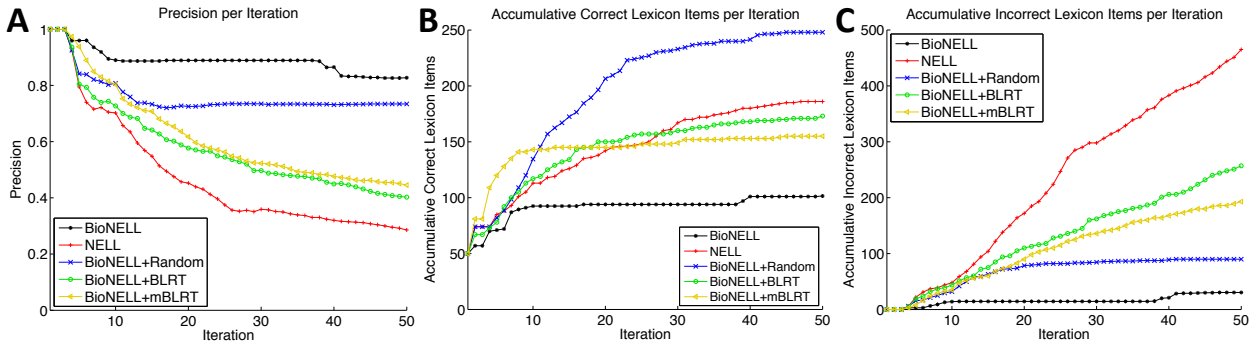

Figure 2: Precision (A), accumulative correct number of lexicon items (B), and accumulative incorrect number of lexicon items (C) per learning iteration for gene lexicons learned using BioNELL and NELL.

One approach for evaluating a set of learned lexicons, the knowledge base, is to select some set of learned instances and assess their correctness [7]. This is a relatively easy task when data is 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. This leads to a possible alternative evaluation approach, 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 thus evaluated lexicons learned by BioNELL by comparing them to available semantic resources. Lexicons of gene names for certain species are available, and the Freebase database [19], an open repository holding data for millions of entities, includes some biomedical concepts. For most biomedical categories, however, complete lexicons are scarce.

4.2.3 Data Sets

To estimate BioNELL's ability in learning lexicons of biomedical categories, we compared the final lexicons learned after 50 iterations, to category *dictionaries*, lists of terms for a concept taken from the following sources, which we consider as a "Gold Standard".

We used three lexicons of biomedical categories taken from the Freebase database [19]: Disease (9420 terms), Chemical Compound (9225 terms), and Drug (3896 terms).

To evaluate gene names we used data from the BioCreative Challenge [20], an evaluation competition focused on annotations of genes and gene products. The data includes a dictionary of genes of the fruit-fly, *Drosophila Melanogaster*. The dictionary specifies a list of gene identifiers, and all possible alternative forms of the gene name, for a total of 7151 terms, which we consider to be the complete dictionary. Figure 1A contains a sample from the fruit-fly gene dictionary.

We used additional data from BioCreative for performing a named-entity recognition task using learned lexicons. The data includes a set of 108 scientific abstracts, manually annotated by BioCreative with gene IDs of fly genes that are discussed in the text. The abstracts may contain the gene ID or any of the gene names. Figure 1B contains an excerpt from one of the abstracts in

Learning System	Precision	Correct	Total
BioNELL	**83**	109	132
NELL	29	186	**651**
BioNELL+Random	73	**248**	338
BioNELL+BLRT	40	173	430
BioNELL+mBLRT	45	155	348
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's bootstrapping methodology 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 $_{\text{by size 132}}$).

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 lexicons of *Chemical Compound* (CC), *Drug*, and *Disease*, learned with BioNELL and NELL. 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 $_{\text{by size}}$).

the data and two IDs of genes that have been annotated as being mentioned in the text.

4.3 Extending Lexicons of Biomedical Categories

4.3.1 Recovering a Closed Category Lexicon

We used BioNELL to learn the lexicon of a closed category, representing the genes of the fruit-fly, *D. Melanogaster*, a long-established "model organism", used to study genetics and developmental biology. We added this new category to the ontology as a descendant of an existing category *Gene*. As potential seeds we used the full dictionary of gene names from the BioCreative data set.

Two samples of genes from the full dictionary of fruit-fly genes 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 lexicons of gene names using BioNELL and the test systems described in Section 4.2.1 (also see Table 2), with initial PMI and random seed sets as shown in Table 1. All systems expanded the initial sets using data from the PubMed biomedical corpus. We measured the precision, total number of instances, and correct instances of the lexicons learned using each system against the full dictionary of genes. Table 3 summarizes the results.

Using a Rank-and-Learn bootstrapping method, initialized with PMI-ranked seeds, significantly improved the precision of BioNELL's learned lexicon over NELL's original bootstrapping method (an increase from 29% for *NELL* to 83% for *BioNELL*). In fact, all the learning systems that used Rank-and-Learn resulted in lexicons with higher precision than *NELL* (83%, 73%, 45% and 40%), which suggests that constraining the bootstrapping process using iterative seed ranking successfully eliminates noisy and ambiguous seeds. Using PMI proves more successful than using the alternative ranking models, BLRT (with 40% precision versus 83% for PMI), and modified-BLRT (with 45% precision).

Since BioNELL's bootstrapping methodology is highly restrictive, it affects the learned lexicon size as well as 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* and *BioNELL+mBLRT* have learned lexicons of similar sizes (338 and 348 terms, respectively), though the precision of *BioNELL+Random* (73%), which uses PMI for ranking, is significantly higher than that of the *mBLRT* alternative (45%).

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 correct items 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 mentions (or at least frequent mentions) of many genes in the dictionary, or (3) some other limitations exist that prevent the learning algorithm from finding additional class examples.

Lexicons learned using BioNELL's PMI ranking methodology show persistently high precision throughout the 50 iterations, even when the process was initiated using random initial seeds (Figure 2A). By the final iteration, all the systems stop accumulating further significant amounts of correct gene instances (Figure 2B). Systems that use PMI-based Rank-and-Learn bootstrapping also stop learning incorrect instances (*BioNELL* and *BioNELL+Random*; Figure 2C). This is in contrast to *NELL* and the BLRT based methods which continue learning incorrect examples.

Interestingly, the highest number of correct gene instances was learned by using Rank-and-Learn bootstrapping with random initial seeds (248 items; *BioNELL+Random*) rather than PMI ranked initial seeds (109 items; *BioNELL*). Both these systems use PMI ranks to determine which learned instances are used during bootstrapping, and the only difference is in the starting set of seeds. While *BioNELL*'s lexicon precision is higher during the entire learning process, in some cases it may be desirable to achieve higher recall at some cost to precision, and these results indicate that doing so may be possible by allowing a more expressive set of initial seeds. However,

Lexicon	Precision	Correct	Total
BioNELL	90	18	20
NELL	2	5	268
BioNELL+Random	3	3	82
BioNELL+BLRT	6	21	307
BioNELL+mBLRT	7	24	272
Complete Dictionary	9	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 manually-filtered lexicon, and lexicons learned with BioNELL and NELL. BioNELL's lexicon achieves the highest precision of all lexicons, and makes more correct predictions than NELL.

there is no guarantee that any single random set will provide the required expressiveness. Note also that *BioNELL+Random* was initiated with the same randomly sampled set of seeds as *NELL*, but due to the more constrained Rank-and-Learn bootstrapping it is able to achieve both higher recall (248 versus 186 correct instances) and higher precision (73% versus 29%).

4.3.2 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 categories are open — new drugs are being developed every year, new diseases are discovered and named, and varied chemical compounds can be created — the Freebase dictionaries are not likely to cover the "complete" current knowledge of 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).

4.4 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. In our evaluation, we use data from the BioCreative challenge (see Section 4.2.3 and

Figure 1), which includes text abstracts and the IDs of genes that appear in each abstract (an example is given in Figure 1B). 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. The annotator takes as input a lexicon of genes (this may be a manually-compiled list of genes or one that was learned with BioNELL). Given sample text, if any of the terms in the lexicon appear in the text, the corresponding gene is predicted by the annotator to be discussed in the text. Since BioCreative's gene annotations are given by gene IDs (see Figure 1B), the annotator emits as output the set of gene IDs of the genes that were predicted for the sample text, based on the input lexicon. For example, for the text in Figure 1B, given a lexicon that contains the word *yellow*, the annotator would predict the gene ID *FBgn0004034*, which is the ID of the *yellow* gene, since the word 'yellow' appears in the text.

We evaluated annotators that were given as input either: the complete fly-genes dictionary, a manually-filtered version of that dictionary (filtering procedure is described below), or lexicons learned using BioNELL and NELL (described in Section 4.3.1). Using these annotators we predicted gene mentions for all text abstracts in the data. We report the average precision (over 108 text abstracts) and number of total and correct predictions of gene IDs, compared with the labeled annotations for each text (Table 5).

Many gene names are shared among multiple gene variants. For example, variants of the *Antennapedia* gene are normally all referred to as *Antennapedia*, or by an alternative name that describes the specific variation (*e.g.*, *Dgua\Antp*, *Dmed\Antp*, and *Dpse\Antp*). A mention of *Antennapedia* in text could refer to any of these. In our precision measurements for all annotators, 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.

Given a complete dictionary of fly genes, it is possible to use it in full for the recognition task. Any gene from the dictionary that is mentioned in the text would be recovered (resulting in high recall for the annotator). However, the full dictionary contains many ambiguous gene names, including short abbreviations, numbers and polysemous gene names such as: Clueless, With and Band (see more examples in Figure 1). These are occasionally used to refer to specific genes, but are mostly used in different contexts. As a result, the ambiguous gene names contribute many false predictions to the complete dictionary annotator, leading to a low precision of 9%.

Some ambiguous terms can be easily removed from the dictionary using filtering rules: for instance, it is easy to remove short abbreviations and numbers. As an example, *section 9* is the name of a gene whose molecular function is currently unknown, and is therefore named after the functional unit to which it belongs, commonly abbreviated simply by the symbol *9*. Naturally, *9* can more commonly appear in text that does not refer to this gene, and thus removing *9* from our lexicon should improve precision without great cost to recall. This filtering approach can eliminate many noisy predictions, although it is not expected to remove polysemous terms which are not easily recognized without more domain knowledge. We filtered the full dictionary by removing one- and two-letter name abbreviations and terms composed only of numbers and non-alphabetical characters, resulting in a filtered dictionary of 6253 terms. Using an annotator over the filtered dictionary, precision has doubled (18%) with some compromise to the number of correct

predictions (138 versus 153 for the full dictionary). However, the overall precision is still quite low, leading to the conclusion that many false predictions remain due to polysemy in gene names.

Using complete or manually refined gene dictionaries for named-entity recognition has been shown before to produce similar high-recall and low-precision results [4].

Finally, we evaluated annotators on fly gene lexicons learned with BioNELL and NELL. *BioNELL*'s lexicon achieved significantly higher precision (90%) than all 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 higher recall (correctly predicted genes) than *NELL*'s lexicon.

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 of this system is based on constrained bootstrapping in which seeds are ranked at every iteration.

The benefits of continuous seed ranking have been demonstrated, showing that using this method leads to significantly less ambiguous lexicons for all the evaluated biomedical concepts. Using BioNELL we see an increase of 51% over NELL, in the precision of a learned lexicon of chemical compounds, and an increase of 45% on a category of gene names. Importantly, when BioNELL and NELL learn lexicons of similar size, BioNELL's lexicons have both higher precision and higher 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 settings, including, a ranking methodology that leverages the current state of the KB, and discovery of relations between ontology categories.

References

[1] H. Ahonen-Myka and A. Doucet. Data mining meets collocations discovery. *Inquiries into words, constraints and contexts*, pages 194–203, 2005.

[2] 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. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

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

[4] R. Bunescu, R. Ge, R.J. Kate, E.M. Marcotte, and R.J. Mooney. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155, 2005.

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

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

[7] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr, and T.M. Mitchell. 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, 2010.

[8] Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010)*, 2010.

[9] B. Carpenter. Phrasal queries with lingpipe and lucene: ad hoc genomics text retrieval. *NIST Special Publication: SP*, pages 500–261, 2004.

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

[11] L. Chen, H. Liu, and C. Friedman. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248, 2005.

[12] K. Church, W. Gale, P. Hanks, and D. Kindle. 6. using statistics in lexical analysis. *Lexical acquisition: exploiting on-line resources to build a lexicon*, page 115, 1991.

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

[14] J.R. Curran, T. Murphy, and B. Scholz. Minimising semantic drift with mutual exclusion bootstrapping. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 172–180. Citeseer, 2007.

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

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

[17] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

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

[19] Google. Freebase data dumps. http://download.freebase.com/datadumps/, 2011.

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

[21] Z. Kozareva and E. Hovy. 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, 2010.

[22] M. Krallinger, A. Valencia, and L. Hirschman. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol*, 9(Suppl 2):S8, 2008.

[23] J. Krishnamurthy and T.M. Mitchell. Which noun phrases denote which concepts? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 570–580. Association for Computational Linguistics, 2011.

[24] A.A. Morgan, L. Hirschman, M. Colosimo, A.S. Yeh, and J.B. Colombe. Gene name identification and normalization using a model organism database. *Journal of Biomedical Informatics*, 37(6):396–410, 2004.

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

[26] P. Pantel, E. Crestan, A. Borkovsky, A.M. Popescu, and V. Vyas. 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, 2009.

[27] E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the National Conference on Artificial Intelligence*, pages 474–479. JOHN WILEY & SONS LTD, 1999.

[28] 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. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 37:5–15, Jan 2009.

[29] L. Tanabe and W.J. Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124, 2002.

[30] Y. Tsuruoka, Y. Tateishi, J.D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392, 2005.

[31] G. Venturi, S. Montemagni, S. Marchi, Y. Sasaki, P. Thompson, J. McNaught, and S. Ananiadou. Bootstrapping a verb lexicon for biomedical information extraction. *Computational Linguistics and Intelligent Text Processing*, pages 137–148, 2009.

[32] V. Vyas, P. Pantel, and E. Crestan. Helping editors choose better seed sets for entity set expansion. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 225–234. ACM, 2009.

[33] R.C. Wang and W.W. Cohen. 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, 2009.

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

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA15213

Carnegie Mellon