# Bootstrapping Biomedical Ontologies for Scientific Text using NELL

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

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

## ABSTRACT

We describe an open information extraction system for biomedical text based on NELL (the Never-Ending Language Learner)[6], 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 biomedical resources. We show that NELL's bootstrapping algorithm is susceptible to ambiguous seeds, which are frequent in the biomedical domain. 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 the original NELL system 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)[6] 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", labeled examples for each ontology category. The new facts, termed *beliefs*, are stored in a growing structured knowledge base.

One of the concerns of 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 quite reliable, as this is ensured by the peer-review process. The 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 researches 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 hence 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: the 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 the wide community of researchers in this field. 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 ontologies 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 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. Redesigning a new ontology for a technical domain is difficult without non-trivial knowledge of the domain. 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. 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, NELL's existing bootstrapping algorithm is highly susceptible to noisy and ambiguous terms. Such ambiguities are common in biomedical terminology (some examples can be seen 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* 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 lead to learning that *green* and *gray* are also genes, and they may not be. In NELL, ambiguity is limited using coupled semi-supervised learning[5]: if two categories in the ontology are declared as mutually exclusive, positive examples of one can be used as negative examples for the other. Thus, to solve the problem of the *white* gene using mutual exclusion, we would have to include a *Color* category somewhere in our ontology, and declare it mutually exclusive with gene names. It is hard to estimate what additional categories should be added, and building a "complete" ontology tree is practically infeasible. It has been shown that biomedical terminology contains a higher rate of ambiguous terms than ordinary English words[10], making this problem a limiting factor in BioNELL.

Recently, NELL has been extended with a method for detecting and compensating for ambiguity — a method which we use in our experiments. 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*, if such categories are present in the ontology[22]. By adding a *Color* category to the ontology, this component can identify that *white* is indeed polysemous. 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, and can potentially make better seeds.

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

**Table 1: Two samples of genes of the fruit-fly *D. Melanogaster*, taken from the complete dictionary of fly genes. *High PMI Seeds* are the top 50 terms from the dictionary selected using PMI ranking, and *Random Seeds* are a random draw of 50 terms. These sets of genes are used as seeds for the *Fly Gene* category (described in full in Section 4.3). Notice that the random set contains many terms that are not distinct gene names including *dad*, and *Bob*. Using these as category seeds can lead the system to learn unwanted beliefs. In contrast, the PMI seeds exhibit much less ambiguity.**

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 of 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. Category names in BioNELL are well-defined by the underlying ontologies and so is their hierarchical relationship. We leverage this fact and rank each seed against a lineage of categories leading to it in the ontology structure. In other words, a *Fly Gene* is also a *Gene* and a *Molecule*, and this information is available through the relationship of these categories in our elaborate ontology.

NELL's bootstrapping algorithm uses the morphological and semantic features of seeds to propose new *beliefs*, which are added to the knowledge base, and used as seeds in the next bootstrapping iteration to learn more beliefs. This means that ambiguous terms may be introduced to 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 beliefs that have been added to the knowledge base by their quality as potential category seeds. Only high-ranking beliefs are added to the collection of seeds that are used in the next bootstrapping iteration. Low-ranking beliefs 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, 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 short 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 give a background description on NELL's bootstrapping algorithm. We also describe the extension of BioNELL using a seed quality collocation measure, and the way in which it is incorporated in the Rank-and-Learn methodology. 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.** Biomedical information extraction systems have traditionally targeted recognition of few distinct biological entities[29], with most focusing mainly on genes and proteins[23, 9, 28, 8]. Few systems (such as the ones described in [33], [15], and more recently [30]) have been developed for fact-extraction of a larger set of biomedical predicates, and these are relatively small scale[33], or they account for limited biomedical subdomains[15] or corpora concerning specific species[30]. We suggest a more general approach, using bootstrapping to extend existing biomedical ontologies, which does not limit possible corpora or predicate selection. The current implementation of BioNELL includes over 100 categories. To the best of our knowledge, large-scale biomedical bootstrapping has not been done before.

**Bootstrap Learning and Semantic Drift.** Carlson *et al.* use a coupled semi-supervised bootstrap learning approach in NELL[6] 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[25]. An ambiguous set of seeds can lead to the problem of "semantic drift", accumulation of erroneous terms and contexts when learning a semantic class[13]. Strict bootstrapping environments reduce this problem by adding boundaries and limitation to the learning process, including learning mutual terms and contexts[26] and using mutual exclusion and negative class examples[13]. Biological terminology, and especially gene names, have been shown to exhibit greater ambiguity than English words[10], suggesting that more aggressive restrictions are necessary in this context to prevent semantic drift. In BioNELL, the initial seeds given to the bootstrapping system are taken from biological, chemical and medical ontologies, 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.* suggest a method for reducing ambiguity in seeds provided by human experts[31], 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 is suggested by Kozareva *et al.*[20] 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 beliefs, 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)[12], the t-test[11], and binomial log-likelihood ratio test (BLRT)[16]. A review of the benefits and short-coming 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.

**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[10]. 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.* provide an in-depth review of the applications of information extraction to biology, and discuss the characteristics of this domain-specific terminology in greater detail[21].

## 3. IMPLEMENTATION

We have implemented BioNELL based on the system design and bootstrapping approach of NELL. In this section we include a short background description of NELL's bootstrapping algorithm (more details on NELL's implementation are presented in [6]). We then describe the data used to build BioNELL, including the text corpus and base ontologies. We describe an automatic process for merging these into one ontology with seed examples for every category. Finally, we define a metric for seed ranking using a PMI collocation measure, and present how this ranking is used in BioNELL in a Rank-and-Learn bootstrapping methodology. We also describe an alternative collocation measure, which we used as a baseline to 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, termed *beliefs*. Three underlying subcomponents operate to suggest candidate facts to the knowledge base. One component extracts free text from the corpus using semantic patterns[7]. The second builds Web queries using currently known facts from the knowledge base, and mines the results for new candidate beliefs[32]. 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, all proposed candidates are examined, and the ones with the most strongly supported evidence are promoted to the status of *beliefs*, and added to the knowledge base. With this process, the KB of beliefs grows with every iteration. This process and all system sub-components are described in greater detail by Carlson *et al.*[6] and Wang and Cohen[32].

At every learning stage, all the beliefs in NELL's knowledge base are used as seeds in the next iteration of learning. This makes the system susceptible to noisy and ambiguous beliefs. As a general approach, ambiguity in NELL is avoided by using mutual exclusion between ontology categories. Beliefs from one category are then used as negative examples for a mutually exclusive category. The use of mutual exclusion relationships in NELL is explained in greater detail in [6].

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

## 3.2 Text Corpora

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 is used by the bootstrapping algorithm to extract beliefs that are added to the knowledge base.

BioNELL's seed-quality collocation measure (described in Section 3.4) is based on a domain-independent Web corpus, the English portion of the ClueWeb09 data set[4], which includes 500 million web documents.

## 3.3 Ontology

BioNELL's ontology includes terms from six base ontologies, covering a wide range of terms from biology, chemistry, and medicine: the Gene Ontology (GO)[2], describing genes and gene product attributes; NCBI Taxonomy for model organisms[27]; Chemical Entities of Biological Interest (ChEBI)[14], a dictionary of molecular entities and small chemical compounds; the Sequence Ontology (SO)[17] for describing biological sequences; the Cell Type Ontology[3]; and the Human Disease Ontology[24].

We used an automatic process for merging the base ontologies into one ontology tree, as follows (also see illustration in Figure 1). First, we group the six ontologies under one hierarchical structure, producing an ontology 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 terms that are unambiguous (have a single parent in the ontology tree), for which we have many potential examples (at least 100 descendants in the sub-tree of the term). This results in 4188 potential categories. 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. In the final ontology tree, only leaf categories are assigned seeds, and these are extended using the learning process. *Potential seeds* are taken from the remaining terms and all synonyms, around 1 million entities, from the full tree. Each populated category is assigned the potential seeds from the sub-tree of the term representing the category. Seed set refinement is described in Section 3.4.

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. Many of these seeds 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 beliefs that are used for learning and expanding a category at any of the bootstrapping steps.

### 3.4.1 PMI Collocation with the Category Name

We consider a category's semantic class as the set of documents that contain a mention of that category. We assess the

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

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

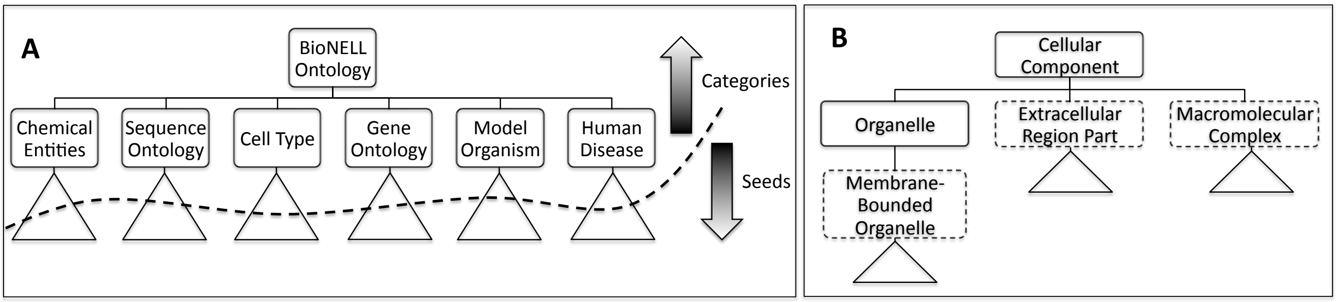

**Figure 1: Illustration of the building process of BioNELL's ontology tree: (A) Six base ontologies are grouped under one root. Then, the terms in the trees of each base ontology (triangles) are divided into a set of categories and seeds; (B) In the final ontology, only leaf categories (dashed rectangles) are populated by the bootstrapping algorithm. The seeds for each category are taken from the sub-tree of the term representing that category (triangles) in the full tree of 1 million terms (the final sub-tree of the category *Cellular Component* is shown).**

quality of seeds by their direct connection to the semantic class, as measured by their collocation with the class name. We avoid selecting ambiguous seed by penalizing those which are frequent in the entire data, relative to their frequency in the constrained set of documents belonging to the class.

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 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[12], 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.

Categories in BioNELL's ontology are part of a hierarchical structure. We leverage this structure to extend seed ranks by measuring the collocation of seeds with their three recent ancestors in the 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. 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 beliefs that have been added to the knowledge base. Only high-ranking beliefs are added to the collection of seeds that are used in the next iteration. Beliefs 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 sub-components 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 belief 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)[16] 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

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

where $k_1 = |Occur(s,D_c)|$, $k_2 = |Occur(s,D)|$, $n_1 = |D_c|$, $n_2 = |D|$, $p_i = \frac{k_i}{n_i}$, $p = \frac{k_1+k_2}{n_1+n_2}$ and

$$L(p,k,n) = p^k(1-p)^{n-k} \quad (4)$$

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 $p_2 >> p_1$) 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 for 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 as well as Rank-and-Learn bootstrapping.

## 4.1 Use-Cases for BioNELL

BioNELL was designed to populate a KB of biomedical categories with facts. 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 beliefs are being introduced by the bootstrapping algorithm. At the end of every iteration, BioNELL contains a lexicon that has been learned for every biomedical concept in the ontology.

A given lexicon for a concept can be used to recognize this concept in free text, for example, using a simple strategy of 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 fact, in our evaluation we show that, for some biomedical concepts, it is better to use an incomplete learned lexicon than a complete one.

## 4.2 Experimental Settings

### 4.2.1 Configurations of the Algorithm

In our experiments, we ran BioNELL using the following configurations of the algorithm (described below and summarized in Table 2), all using the biomedical corpus and the ontology described in Sections 3.2 and 3.3. The system ran for 50 iterations under all configurations, in order to evaluate the long term effects of ranking on the knowledge base. Under each system configuration we distinguish a test category for which we assess the quality of the beliefs predicted by the system, comparing it against a Gold Standard dictionary (data for these is described in Section 4.2.3). 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 one of two bootstrapping methods: (1) BioNELL's Rank-and-Learn bootstrapping (described in Section 3.4.2), and (2) NELL's bootstrapping algorithm (see Section 3.1 and [6] for more details). In each of those configurations, we used one of two possible sets of 50 initial seeds: (1) the top 50 seeds using PMI ranking with the category name, and (2) a random set of seeds taken from the category's potential seeds. As a baseline to the PMI ranking model, we used two additional configurations using BioNELL's bootstrapping methodology where PMI ranks were replaced with BLRT and modified-BLRT ranks (described in Section 3.4.3). Table 2 contains a succinct summary of all configurations.

| Learning System Configuration | Bootstrapping Algorithm | Initial Seeds |
|---|---|---|
| **BioNELL** | Rank-and-Learn with PMI | PMI top 50 |
| **BioNELL+Random** | Rank-and-Learn with PMI | Random 50 |
| **NELL** | NELL's algorithm | PMI top 50 |
| **NELL+Random** | NELL's algorithm | Random 50 |
| **BioNELL+BLRT** | Rank-and-Learn with BLRT | BLRT top 50 |
| **BioNELL+mBLRT** | Rank-and-Learn with mBLRT | mBLRT top 50 |

**Table 2: Learning system configurations used in the evaluation, including the main configuration *BioNELL*, and five baseline configurations used for testing the ranking and bootstrapping approach used in the main configuration.**

### 4.2.2 Evaluation Methodology

Using BioNELL we can learn *lexicons*, collections of terms, for categories in the ontology. The *lexicon* is the collection of instance names that were learned for a category after using the system.

One approach for evaluating a set of learned lexicons, the knowledge base, is to select some set of beliefs from the knowledge base and assess their correctness[6]. This is a relatively easy task when data is extracted for general categories like City or Sports Team. For example, it is easy to say that the statement "London is a City" is correct. 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, as most people are not necessarily familiar with the details of the periodic table. 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 any 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. For example, lexicons of gene names for certain species are available, and the Freebase database[18], an open repository holding data for millions of entities, includes several biomedical concepts. For most biomedical categories, however, complete lexicons are scarce. We evaluated three categories from our ontology for which we found corresponding dictionaries in Freebase, and we extended the ontology with an additional category, evaluated with data from the BioCreative challenge.

#### 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[18]: Disease (9420 terms), Chemical Compound (9225 terms), and Drug (3896 terms).

To evaluate gene names we used data from the BioCreative Challenge[19], an evaluation competition focused on annotations of genes and gene products. The data includes a complete dictionary of genes of the fruit-fly, *Drosophila Melanogaster*. The dictionary specifies a list of gene identifiers, including the common name for each gene and all possible alternative forms of the gene name, a total of 7151 terms.

We used additional data from BioCreative for performing a named-entity recognition task using BioNELL's lexicons. The data includes a set of 108 scientific abstracts, manually annotated by BioCreative with gene identifiers for genes of the fruit-fly that are discussed in the text. The abstracts may contain the common gene name or any of the alternative forms.

## 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 category to the ontology as a descendant of an existing category *Gene*. As potential seeds to this new category we used the full dictionary of gene names, taken from BioCreative.

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 *dad*, and *Bob*. In contrast, the PMI seeds exhibit much less ambiguity. We used BioNELL to learn lexicons of genes using the system configurations described in Section 4.2.1 (also see Table 2), with the seed sets shown in Table 1. We measured precision and recall of the lexicons learned from each learning system against the full dictionary of genes. Table 3 summarizes the comparison results.

| Learning System | Precision | Lexicon Size |
|---|---|---|
| BioNELL | **83** | 132 |
| BioNELL+Random | 73 | 338 |
| NELL | 38 | **1049** |
| NELL+Random | 29 | 651 |
| BioNELL+BLRT | 40 | 430 |
| BioNELL+mBLRT | 45 | 348 |

**Table 3: Precision and lexicon size of lexicons of fly genes, learned using BioNELL, and compared against the full dictionary.**

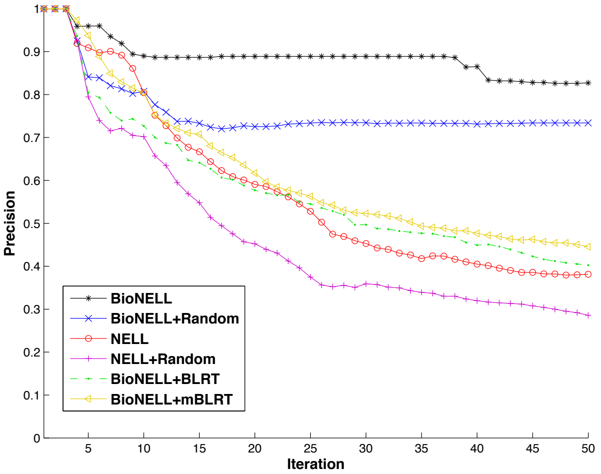

**Figure 2: Precision of gene lexicons over bootstrapping iterations.**

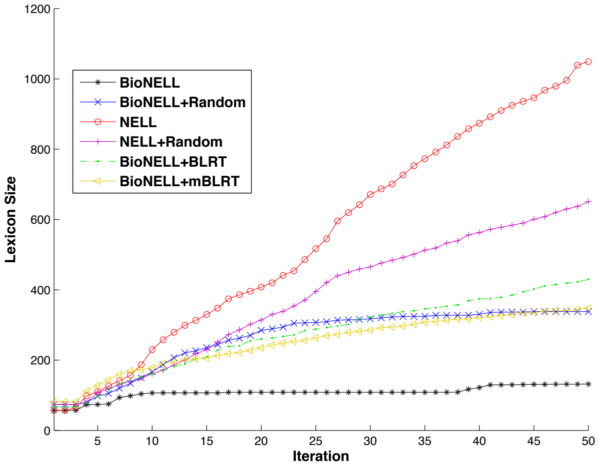

**Figure 3: Size of gene lexicons over bootstrapping iterations.**

Using the Rank-and-Learn methodology, seed refinement using PMI ranking at every learning iteration, significantly im-

| Learning System | Precision | | | Lexicon Size | | |
|---|---|---|---|---|---|---|
| | Chemical Compound | Drug | Disease | Chemical Compound | Drug | Disease |
| BioNELL | **66** | **52** | **43** | 96 | 972 | 624 |
| NELL | 15 | 40 | 37 | **449** | **1300** | **782** |

**Table 4: Precision and lexicon size of lexicons of three open categories, *Chemical Compound*, *Drug*, and *Disease*, learned with BioNELL and NELL.**

proved the precision of the learned lexicons, whether starting with ranked initial seeds (an increase from 38% to 83%), or random seeds (an increase from 29% to 73%). PMI ranking also had a positive effect when considering the initial set of seed: when Rank-and-Learn bootstrapping was used, ranking the initial seeds lead to an increase of 10% in precision (from 73% to 83%), and when NELL's bootstrapping was used, there was an increase of 9% (from 29% to 38%). Using PMI for ranking proves more successful then using the alternative ranking models, BLRT (with 40% precision versus 83% for PMI), and modified-BLRT (with 45% precision).

Despite running for 50 iterations, all the lexicons that have been learned cover a very small portion of the full set of genes (under 6% recall), suggesting either that, (1) more learning iterations are required, (2) the biomedical corpus we use is too small and does not contain documents with mentions of all the 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 methodology show persistently high precision throughout learning iterations, even when the process was initiated using random initial seeds (Figure 2). Since BioNELL's bootstrapping methodology is highly restrictive, it affects the size of the learned lexicon as well (Figure 3). Notice, however, that while the *BioNELL+mBLRT* learning system has learned a lexicon similar in size to the *BioNELL+Random* configuration (the final lexicons include 348 and 338 terms, respectively), the precision of *BioNELL+Random* (73%), which uses PMI for ranking, is significantly higher than that of the *mBLRT* alternative (45%).

### *4.3.2 Extending Lexicons of Open Categories*

We evaluated learned lexicons for three open categories, Chemical Compound, 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.

Two types of system configurations were used to extend lexicons for these categories, both starting with PMI-ranked initial seeds. While recall is similar when using both systems (1% for *Chemical Compound*, 13% for *Drug*, and 3% for *Disease*, for both systems), precision is higher for all three categories when ranking is added to the bootstrapping process (*BioNELL* configuration, Table 4). Similar to the case of learning a closed lexicon, the additional restrictions of the Rank-and-Learn methodology result in smaller sized learned lexicons for open categories as well.

| Lexicon | Precision | Recall |
|---|---|---|
| Complete Dictionary | 9 | **68** |
| Filtered Dictionary | **15** | 63 |
| BioNELL | **90** | 8 |
| BioNELL+Random | 3 | 2 |
| NELL | 19 | **13** |
| NELL+Random | 2 | 4 |
| BioNELL+BLRT | 6 | 10 |
| BioNELL+mBLRT | 7 | 12 |

**Table 5: Named-entity recognition using a complete lexicon, a manually-filtered version of the complete lexicon, and lexicons learned using BioNELL.**

## 4.4 Named-Entity Recognition using a Learned Lexicon

We evaluated the use of lexicons learned with BioNELL 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 show that when recognizing gene names, using a "filtered" dictionary, like the one that is learned with BioNELL, is better than using the complete dictionary of genes. The evaluation is based on text abstracts annotated with gene identifiers of genes of the fruit-fly that are mentioned in the text (see Section 4.2.3 for more details on the BioCreative data).

Given a lexicon, we implemented an *annotator* for predicting what genes are discussed in text. A gene is predicted to be mentioned in the text if a term from the lexicon appears in the text, and the term is the gene name, or one of the alternative name forms for that gene. For each abstract, we aggregate the set of gene identifiers of all genes predicted to be mentioned in it. We evaluate annotators, by measuring the precision and recall of the predicted set of gene identifiers, compared with the labeled annotations for each text. We report the average precision and recall over all text abstracts on which we predicted an annotator.

Many gene names are shared among multiple genes. For example, the various mutants of the *Antennapedia* gene are all referred to by the gene common name, or by an alternative name that describes the specific mutation. A mention of *Antennapedia* in the text may refer to any of the mutant forms or the wild-type gene. In our precision measurement for all annotators, we consider a prediction of a gene identifier as "true" if it is labeled as such by BioCreative, or if it shares a synonym name form with another true labeled gene identifier.

First, we evaluated an annotator over the complete fly-genes dictionary, and a manually-filtered version of that dictionary (filtering procedure is described below). Next, we evaluated annotators on lexicons learned using BioNELL (see Section 4.2 and Table 2 for description of learning system configurations). Table 5 summarizes the performance of all the evaluated annotators, and the results are discussed in detail in the following text.

### 4.4.1 Using a Complete Dictionary

One approach to this task is to use the full dictionary of gene names as a lexicon. Names that appear in the text in the same format that is included in the dictionary would all be recovered (resulting in high recall for the annotator). However, the full dictionary of fruit-fly genes contains ambiguous alternative name forms, including single letter abbreviations, isolated numbers and polysemous gene names such as: Clueless, Homeless and Balloon. These are occasionally used to refer to genes, but mostly they are used in different semantic context. As a result, using the full dictionary for this task we get an annotator with very low precision, 9% (Table 5).

Note that using the full dictionary results in recall of only 68%. This is due mainly to some inaccuracies in the annotation data. The text paragraphs in our data are abstracts of articles concerning the fruit fly. In some cases the labeled annotations of an abstract include gene identifiers of genes that are not directly mentioned in the abstract, but rather in the full text of the article, which is not available to us.

### 4.4.2 Using a Manually-Filtered Dictionary

Another possible approach is to remove likely-to-be-ambiguous terms from the full dictionary of gene names using simple filtering rules. This can eliminate some of the noisy predictions, while not handling polysemous terms that are not easy to recognize without specific domain knowledge. We filtered the full dictionary by removing terms, including terms that are composed only of numbers and other non-alphabetical characters, and one- and two-letter name abbreviations. The final filtered dictionary contains 6301 terms. Using an annotator over the filtered dictionary, precision has nearly doubled (15%) without much compromise to recall (63%, Table 5). However, the overall precision is still low, leading to the conclusion that many false predictions are still due to polysemy in gene names.

### 4.4.3 Learning a Lexicon Using BioNELL

We used BioNELL to automatically learn lexicons of fly genes, and evaluated annotators on the learned lexicons. The lexicon learned using the full BioNELL approach, including PMI ranking at every iteration, generates highly accurate predictions, with 90% precision, which is significantly higher than the precision of all other lexicons, including those based on the complete and filtered dictionaries (Table 5). The number of true predictions is low for all learned lexicons (under 13% recall). This could potentially improve with more learning iterations.

## 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 a constrained bootstrapping approach where seeds are ranked at every iteration.

The benefits of using continuous seed ranking have been demonstrated, showing a significant decrease in ambiguity in learned lexicons for 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. BioNELL's gene lexicon substantially outperforms all alternative lexicons, when used for 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, a model for distinguishing well-known from novel facts, and discovery of relations between ontology categories.

## 6. 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.

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

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

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

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

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka Jr, and T. 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.

[7] A. Carlson, J. Betteridge, R. C. Wang, E. R. H. Jr., and T. 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.

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

[9] J. Chang, H. Sch

”utze, and R. Altman. Gapscore: finding gene and protein names one word at a time. *Bioinformatics*, 20(2):216, 2004.

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

[11] 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.

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

[13] J. 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.

[14] 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.

[15] 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.

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

[17] K. Eilbeck, S. Lewis, C. 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] J. Krishnamurthy and T. 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.

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

[24] 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.

[25] P. Pantel, E. Crestan, A. Borkovsky, A. 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.

[26] 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.

[27] 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.

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

[29] Y. Tsuruoka, Y. Tateishi, J. 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.

[30] 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.

[31] 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.

[32] R. Wang and 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.

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