

ACADEMIC
PRESSAvailable online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Biomedical Informatics 35 (2002) 247–259

Journal of
Biomedical
Informaticswww.elsevier.com/locate/yjbin

Rutabaga by any other name: extracting biological names

Lynette Hirschman,* Alexander A. Morgan, and Alexander S. Yeh

The MITRE Corporation, MS K312, 202 Burlington Rd., Bedford, MA 01730, USA

Received 10 September 2002

Abstract

As the pace of biological research accelerates, biologists are becoming increasingly reliant on computers to manage the information explosion. Biologists communicate their research findings by relying on precise biological terms; these terms then provide indices into the literature and across the growing number of biological databases. This article examines emerging techniques to access biological resources through extraction of entity names and relations among them. Information extraction has been an active area of research in natural language processing and there are promising results for information extraction applied to news stories, e.g., balanced precision and recall in the 93–95% range for identifying person, organization and location names. But these results do not seem to transfer directly to biological names, where results remain in the 75–80% range. Multiple factors may be involved, including absence of shared training and test sets for rigorous measures of progress, lack of annotated training data specific to biological tasks, pervasive ambiguity of terms, frequent introduction of new terms, and a mismatch between evaluation tasks as defined for news and real biological problems. We present evidence from a simple lexical matching exercise that illustrates some specific problems encountered when identifying biological names. We conclude by outlining a research agenda to raise performance of named entity tagging to a level where it can be used to perform tasks of biological importance.

© 2003 Elsevier Science (USA). All rights reserved.

1. Background

With the successful sequencing of the human genome, the amount of published literature is growing exponentially, including both journal articles and biological databases. This explosion is causing biologists to turn increasingly to information technology to organize, access, and process the information [1].

Biological information takes different forms. Many model organisms are now being sequenced;¹ however, sequencing and gene identification are increasingly taken for granted. Sequences have become the raw material which must be “mined” for information and the focus of research has moved to understanding the

functions of the genes and the pathways that regulate the expression of the genes. Because of the rapid proliferation of data, researchers have created, by hand, many special purpose databases to provide them with convenient access to specific kinds of information. Some databases are organized around a specific organism (Flybase: <http://www.flybase.org>; Yeast: <http://genome-www.stanford.edu/Saccharomyces/>, etc.). Other databases specialize in genes, proteins, protein structure, pathways, or gene expression. There are now over 280 specialized databases of biological data.² Most of these databases are created by labor-intensive expert “curation” of the entries: Ph.D. biologists read the literature and transfer the key pieces of information from the journal articles into the appropriate fields in the databases, using a controlled vocabulary or, in some cases,

* Corresponding author. Fax: 1-781-271-2352.

E-mail address: lynette@mitre.org (L. Hirschman).

¹ See <http://www.nih.gov/science/models/activities/index.html> for a list of organisms now being sequenced, including include the puffer fish, chicken, sea urchin; the rice genome was recently published, along with mouse; and TIGR has recently announced that it will sequence entire ecosystems.

² See the list of biological databases at <http://www.infobiogen.fr/services/dbcat>. This list includes 511 databases, although some databases appear in multiple tables.

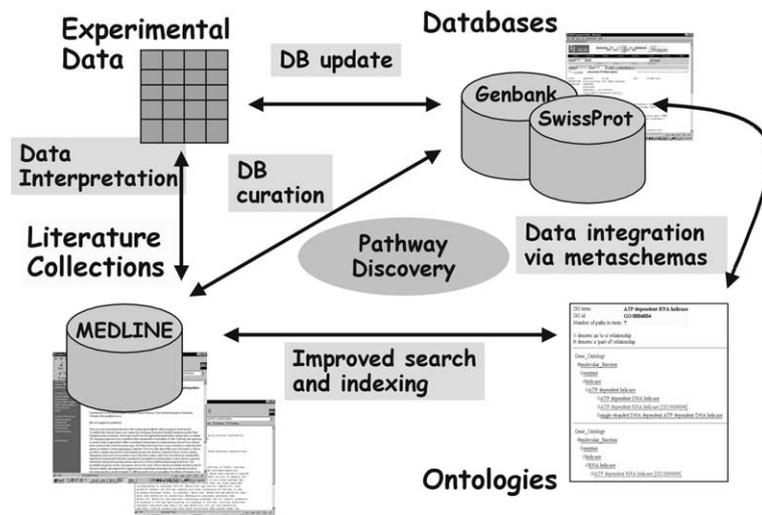


Fig. 1. Applications for information extraction in managing biological information.

an ontology (e.g., the Gene Ontology or GO³). These databases were originally designed for direct access by biology researchers; today, with the increasing volume of information and the proliferation of databases, biologists access the information indirectly through advanced interfaces that perform complex searches and aggregate information from multiple databases and other information resources.

1.1. Why names are important

Names of biological entities provide the critical links across these different information sources—we can think of them as *indices* into the data. This article focuses on the use of biological names⁴ for information access, looking specifically at why this is a hard problem and what techniques are being developed to assist biologists. Research in natural language processing (NLP) has developed a number of techniques that are now being applied to information access in biology. If there were techniques to automatically identify and normalize names of biological entities and relations among these entities, then there are many places where these techniques could be applied.

Fig. 1 shows potential applications for information extraction. In the four corners are sources of data: experimental (structured) data, curated databases (semi-structured data), ontologies, and the literature collec-

tions. Linking these are information access and language processing techniques.

Database curation aids provide a (semi-automated) mapping from literature to database, as evaluated in the recent Challenge Cup Evaluation for Knowledge Discovery and Data Mining Conference 2002 (see <http://www.biostat.wisc.edu/~craven/kddcup/tasks.html>).

Ontology mapping techniques can be used to provide meta-data for improved literature search and indexing across multiple databases [2]. Ontologies can be linked to the literature using information extraction techniques to add new terms to the ontology [3] or to associate ontology terms with mentions in the literature [4].

Data mining can improve interpretation of experimental data, including both BLAST search [5] and interpretation of micro-array data [6]. And text data mining can be used for discovery of new relations, e.g., pathways [7,8] shown in the center of Fig. 1.

All of these tools depend on correct biological entity identification—that is, extracting the biological entities mentioned in each data source and mapping them to a canonical form, so that they can be used to cross-index the information.

1.2. Extracting names

Biological names are complex. The language used to describe biological knowledge is constantly changing because our understanding of biology constantly expands. New terms are added to name the new entities (genes, proteins, and pathways); in some cases, terms are removed or refined as knowledge is reorganized. The names are used to communicate knowledge among researchers. This means that biology is particularly dependent on shared naming conventions. If biologists cannot make a unique mapping from name to underly-

³ The Gene Ontology (GO) can be found at the Gene Ontology Consortium web page, <http://www.geneontology.org/>; GO produces a controlled vocabulary to describe molecular function, biological process and cellular component.

⁴ We restrict our attention here to terminology related to genomics: genes, proteins, gene expression products, and sub-cellular localization. For different subfields of biology or medicine, the terminology may be quite different and will likely present somewhat different challenges for information extraction.

ing object (gene or protein or structure), then there is possible failure of communication.⁵ Scientific progress is slowed when researchers can no longer build on each other's experiments, leading to needlessly duplicated research.

Building reliable natural language processing tools for biology is critical to managing biological information. Recent research in natural language processing has produced a number of methods and even some commercial tools that improve information access in non-biological domains (particularly, for news). Section 2 looks at these results for automatic name and relation extraction, reviewing the results for news stories and the corresponding results in biology. In Section 3, we discuss why named entity extraction is difficult, with particular reference to problems in biology. Section 4 looks at issues specific to naming in genomics and describes a simple experiment using lexical resources to identify biological terms. The results from this experiment highlight both the resources available for biological name extraction as well as some of the problems specific to identifying biological names. Finally, Section 5 concludes with a research agenda to boost the performance of biological name extraction to levels that will support performance of useful biological tasks.

2. Extracting names in biology

This section describes information extraction first for news, then for biology. We review the series of evaluations done in the context of the Message Understanding Conferences (MUCs), and similar kinds of experiments for extraction of names and relations in biology.

Recent research in NLP has developed technologies capable of automating various aspects of information processing. We can divide the technology into three broad areas:

Information retrieval. Retrieval of documents in response to a query or list of key words. This is the technology that underlies modern search engines. In the medical domain, the most familiar example is PubMed, which supports retrieval of on-line biomedical abstracts with rich links to related resources. Information retrieval algorithms rely on term co-occurrence techniques for clustering and classifying documents. These shallow, non-linguistic techniques are computationally efficient and can index huge collections and return sets of appropriate documents quickly. They provide a trade-off between precision and recall, so that it is possible to

retrieve a few relevant documents with high precision, but exhaustive retrieval (high recall) typically requires sifting through many irrelevant documents.⁶

Information extraction. Extraction of names and entities, relations and key facts from text. Extraction may take the form of in-line annotation of text (e.g., to highlight key terms) or it may take the form of lists of entities (a “cast of characters”) or tables of relations. These lists or tables can be used to provide indices into the data, as in a curated database, which indexes information by gene or protein and provides pointers back to the literature from which the information was derived. There are now commercial systems for named entity extraction for news (e.g., person, organization, location, and certain kinds of numerical expressions).⁷ We explore the applications to biology in greater detail below.

Question answering. Return of a phrase or sentence or summary in response to a factual question (in English, for example). Question answering differs from information retrieval in that information retrieval returns lists of documents for the person to read; question answering returns facts or answers, with pointers back to the underlying documents drawn from a large collection of documents. This is a relatively recent area of research and results are promising: systems can return correct answers for 75–85% of simple factual questions in a general domain [9]. However, these techniques have yet to be explored in biology.

2.1. Information extraction for news

Our focus here is on information extraction. Much of the work in information extraction has been focused on news stories, in the context of a series of seven Message Understanding Conferences [10]. These started in 1987 and ended in 1998 with MUC-7. The original focus was on event or Scenario Template extraction. For MUC-6 (1995) [11] and MUC-7 (1998) [12], new intermediate tasks were defined [13], including tasks for identification of Named Entity, Template Element, and Template Relation, described below. These tasks and the associated evaluations were designed as technology-focused exercises to improve the basic technology, although they represent an abstraction of the needs of analysts reading news reports.

⁵ The term MAP exemplifies the possible confusions. This is a word of non-technical English with a specific meaning in genomics; it also has been used to stand for microtubule-associated protein, microsomal aminopeptidase, methionine aminopeptidase, and mitogen-activated protein (as in MAP kinase).

⁶ Precision is defined as the number of correct system responses divided by total number of system responses. Recall is defined as the number of correct system responses divided by the total number of correct responses possible.

⁷ Commercial products include BBN's Identifinder (<http://www.bbn.com/speech/identifinder.html>), IBM's Intelligent Miner for Text (<http://www3.ibm.com/software/data/iminer/fortext/index.html>), and Inxight's Thingfinder (<http://www.inxight.com/products/thingfinder/>).

Named entity consists of the extraction of *named entities in running text*. The task definition for MUC-6 and MUC-7 requires the identification and classification of strings representing proper names of persons, organizations, locations, and, for MUC-7, artifacts, e.g., manufactured objects such as cars or airplanes. It also includes identification of numerical expressions for time, date, money, and percent. As formulated, the task requires no “normalization” of the names. That is, if there is a mention of “General Motors” and another mention of “GM,” these are independently annotated as mentions of organizations, with no explicit relation between them. Typical metrics for named entity are precision and recall, evaluated against a fully annotated “key” or “gold standard.”

Template element consists of the extraction of the *list of unique entities* mentioned in each story or document, along with some associated properties. For example, names of organization discussed in a story would be extracted, including their semantic type and subtype (ORGANIZATION/GOVERNMENT), other name variants or “aliases” and a short descriptive phrase. The template element task is a document level task: the list of template elements is associated with the entire document. It requires aggregating information across the document and resolving coreference,⁸ in order to collapse all mentions into a single template element. Thus for the template element task, “General Motors” might be the organization name, while “GM” would be included as an alias (alternate name). Listing these as separate entities would be an error of commission (or a precision error)⁹ for the template element task.

Template relation consists of extraction of a *specific set of binary relations among entities*. For MUC-7, these included the relations among the major categories of person, organization and location, specifically LOCATION_OF, PRODUCT_OF, EMPLOYEE_OF. These relations are captured at the document level, although it is often possible to find the information in a single sentence.

Scenario template is a complex template task capturing a specific kind of event, e.g., corporate succession, or a satellite launch. The scenario template has a nested structure consisting of embedded templates for relations, with slots for temporal and geo-spatial information as well as for status (past, future, or planned). Building the scenario template requires integrating in-

⁸ Resolving coreference means associating successive mentions of the same entity in the text with a single identifier. Coreference may take the form of a synonym (General Motors. . . GM) or reference by pronoun (GM. . . It is located in Detroit. . .), or reference by a noun phrase, as in “GM. . . The company is located in Detroit.”)

⁹ Returning an incorrect response or an extra response is an error of commission, or a precision error; it is also referred to as a false positive. Failing to return an entity is an error of omission or a recall error. This is referred to as a false negative.

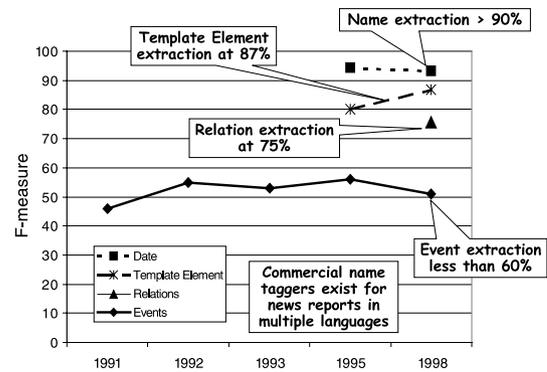


Fig. 2. Performance on MUC tasks over time.

formation found throughout the article into the single template. This was the original MUC task and, as noted below, it proved quite difficult.

The MUCs provided a common task definition or challenge evaluation, plus training and test sets (along with evaluation software) for each evaluation. The series of evaluations made it possible to measure the progress on these tasks over time across a research community. Fig. 2 plots progress over time; the *x*-axis shows the different evaluations labeled by year; the *y*-axis is the harmonic mean of precision and recall, called *F*-measure.¹⁰ The points for each task in a given year represent the results from the highest performing system for that task.

We see from this figure that the named entity task is the easiest, with systems scoring in the mid-90’s for English [14]. The results for Japanese (not shown) were almost as good at 91% *F*-measure [15]; the systems for Chinese showed significant improvement, reaching an *F*-measure of 86% by 1998 [16]. For named entity, systems generally use local information to identify multi-word expressions in context. Successful approaches have often included part-of-speech tagging (a sequence of proper nouns as a name), selection of appropriate word features (particularly Case, but other features for times and dates), learning of local context cues (e.g., “Mr.” often precedes a person name), and use of specialized lexicons and gazetteers (for place names).

The template element task also showed good results by MUC-7 in 1998, with the best system receiving an *F*-measure of 87% [17]. This is a document level task and requires the ability to recognize synonyms and multiple mentions of the same entity, in order to avoid duplicate entries. Relation extraction also showed promise, with an *F*-measure of 75% by 1998 [17]. This is a more complex task, requiring that the system be able to capture predicate–argument structures. While mentions of a

¹⁰ *F*-measure is the harmonic mean of precision (*P*) and recall (*R*): $F = 2 * P * R / (P + R)$.

relation frequently occur within a single sentence or even within a phrase, in other cases, the information may be spread across multiple sentences. Finally, the scenario template scores for extracting complex events has remained largely unchanged, staying in the 50–60% *F*-measure range [17]. This task requires not only entity coreference but event coreference as well, since all information about an event must be captured in one scenario template.

These results tell us several things. First, some of the extraction tasks (named entity and template element) have reached good precision and recall values for news stories. Second, it illustrates how the creation of a common evaluation was able to attract a number of research groups to work on a particular problem. A total of 16 groups participated in MUC-6, and the same number three years later in MUC-7. Based on these tests repeated over time, it has been possible to bring a large research community together to determine what techniques work well and how to engineer systems for improved performance on particular tasks. For multilingual named entity tagging, there has been significant progress, resulting in several commercial systems for name tagging. To date, these systems are limited to variants of the basic set developed for news, identifying persons, organizations, and locations plus associated attributes, such as title, address, phone number, etc.

A number of factors have contributed to these successes. First, the repeated evaluations brought a significant set of intellectual resources to bear on this problem. Second, the availability of well defined tasks and annotated training sets made it possible to experiment with a variety of machine learning and statistical techniques. Statistical techniques, in particular, Hidden Markov Models (HMMs), were applied successfully to named entity tagging, e.g., the Nymble system [18].

A weakness of the tasks as defined in MUC is that the technology was developed independent of any specific application. The original scenario template task was an abstraction of a real problem, namely the consistent filling of a database with events or relations of interest to analysts. However, there were no real databases that were used to provide training and test data, and no actual users. As a result, all the materials had to be developed specifically for the MUC evaluations, by hand. This was a labor-intensive process that restricted the amount of training data available. For the named entity task, the amounts of training data seemed to be sufficient, although subsequent work has shown that the error rate continues to drop as a function of the log of the amount of data [18]. However, for the document level tasks (particularly template relation and scenario template), the density of relations or events is lower, there is more linguistic variability and the requirement

for training data is therefore higher. This may be one explanation for the apparent plateau in scenario template results as seen in Fig. 2. In addition, there were no users to determine what level of performance was good enough to be useful; it was unclear whether the systems should optimize for higher recall (at the expense of lower precision), or focus on techniques for high precision results. As we will see below, this contrasts with the biomedical domain, where users are driving the technology requirements, and there is the potential for a much tighter coupling between technology and application.

2.2. Information extraction in biology

Given the encouraging results for named entity, template element, and relation extraction applied to news, it would seem plausible to expect the same performance figures for biology. Since the late 1990s, biologists and computer scientists have been applying both natural language and co-occurrence based text data mining techniques to biology to address the applications parallel to those for news. However, what we see (Table 1) is that after relatively high scores reported in several early experiments on name extraction for small data sets, the results on larger data sets seem to cluster in the 75–80% range. Surprisingly, the results for relation extraction seem more comparable to news (75% for news, 65–75% for biology).

The tasks for extraction from news have their obvious counterparts in biology. For example, *named entity extraction* for genomics would include extracting names of genes, proteins, small molecules, sub-cellular localization, tissue type, organism, etc. The *relations* of interest include activation, signaling, reaction or, at a higher level, general gene–protein, protein–protein, and transcript–protein interactions. There have been an increasing number of results reported in these areas; however, most of these experiments have been done on specialized data sets, making it difficult to compare results across methods. We review the results for named entity and template relations for biology; Table 1 provides an overview of some of experiments, including estimated corpus size (training and test sets), number of classes, and reported precision, recall and *F*-measure.

For named entity tagging, initial results seemed quite good. Fukuda et al. [19] used a very small test set and hand-crafted rules to identify protein names. They reported results of 91.9% precision and 93.3% recall on a set of 30 abstracts.¹¹ The system by Proux et al. [20] reported comparable results: 94.4% recall and 91.4%

¹¹ There are higher results reported in [19]; however, it is not clear whether these higher results (precision 94.7%, recall 98.8%) were obtained on blind data.

Table 1
Comparative Named Entity results, including human interannotator agreement results

Named entity	Ref.	Author	Date	Data source	# Wds (estim)	# Class	P	R	F
Biology	[19]	Fukuda	1998	MEDLINE	20,000	2	92	93	93
	[20]	Proux	1998	Flybase summary	12,000	1	91	94	92
	[22]	Collier	2000	MEDLINE	30,000	4			73
	[21]	Krauthammer	2000	Review article	5000	1	72	79	75
	[23]	Gaizauskas	2002	MEDLINE	30,000	12	84	82	83
	[23]	Gaizauskas	2002	MEDLINE	6000	12	92	86	89
	[21]	Krauthammer	2000	Review article	5000	1	93	76	84
News wire	[18]	Bikel	1999	News English	650,000	6	93	96	95
	[13]	Chinchor	1998	English	50,000	6	98	96	97
		human annotators							

precision (balanced F measure of 92.5%). Their experiment made use of 1200 sentences extracted from Flybase which were known to contain at least two gene symbols (simple one-word gene names). The system used a tokenizer and tagger, followed by a lexical stage of error recovery and then contextual analysis. The lexical stage included lexicons of English and biological terms, as well as prefixes and suffixes for English and biological terms. These results are not comparable to the named entity extraction results from MUC, because the data set was carefully selected to contain positive examples, and because only single word expressions of one class were captured.

After these initial results, more recent results of named-entity-like tasks have ranged from 73% to 83%. In a novel approach, Krauthammer et al. [21] encoded gene names and text in terms of DNA 4-tuples and used the BLAST algorithm to look for “homologies” between the text and known gene names. This innovative method achieved results of 78.8% recall and 71.7% precision for gene names. Interestingly, this method required no training data; it relied instead on similarities between gene names listed in GenBank and those strings appearing in the text.

Collier et al. [22] applied a linear interpolating Hidden Markov Model to extract terminology from MEDLINE abstracts. Using a small (30,000 words) corpus of MEDLINE abstracts (100 abstracts, 80 for training, and 20 for test), they report an F -measure of 72.8% for tagging names of proteins, DNA, RNA, and seven kinds of source information (cell line, cell type, sublocation, tissue, etc.).

Recent results reported from the PASTA system [23] are somewhat higher than previous work. Gaizauskas et al. report a recall of 82% and precision of 84% (F -measure of 83%) for the task of identifying 12 classes of entities for information extraction about the roles of residues in protein molecules. The corpus was drawn from MEDLINE abstracts; 133 were used for the named

entity work, including a set for development (52), interannotator agreement studies (20) and a blind test set (61). The system architecture of PASTA used a sequence of processors for section analysis, tokenization and part-of-speech tagging. This was followed by terminological processing, which consisted of morphological analysis, lexical lookup in a compendium of information compiled from biological databases, and finally, terminology parsing, to identify multi-word phrases. The two unusual features of PASTA are its use of morphology tailored to the biomedical domain and a large-scale biological lexicon.

Looking at biological analogues to other MUC tasks, there have been no results for the template element task, but there has been considerable work on extracting relations. Recent results on relation extraction, using MUC-style metrics, have been promising. This is surprising, since relation extraction would appear to be dependent on reliable entity extraction. Let us assume that correct extraction of a binary relation requires identification of three terms—the two participating entities and the specific relation that holds between them. If we estimate performance of named entity extraction at approximately 92% for news and use this figure for the correct identification of the relational term as well, then a simple model that assumes independence of these three subtasks would estimate performance as $0.92 \times 0.92 \times 0.92$ or around 78% F -measure. This is close to the observed value of 75% for news stories. This model also explains the difficulty of complex event extraction because events typically require correct fill of five or six slots, which can degrade performance quite quickly.

However, when we turn to biology, we see an anomaly: named entity extraction hovers in the 75–80% range, but relation extraction seems to also be in the same range. For example, Friedman et al. [24] report results for identifying a wide range of within-sentence relations at a precision of 96% and a recall of 63% (for

an F -measure of 76%).¹² And Gaizauskas et al. [23] report a precision of 65% and a recall of 68% (F -measure of 66%) for MUC-style relations. Both of these results are much higher than would be predicted on the basis of the named entity tagging results. This suggests that the independence assumption in the simple model does not hold for biology relations: the context provided by the relation word (e.g., “enhances” or “phosphorylates”) may provide strong cues for the occurrence of gene or protein terms as arguments, improving the named entity tagging performance in the vicinity of such verbs.

3. Are names in biology harder than names in news?

These results raise the question: why do results for biology appear to be worse than for news? There are a number of possible explanations. First, we do not really know if the results are worse, since there have been no common evaluations in biology. It is unclear how to compare biology name tagging systems to each other, and even more unclear how to compare results across domains. However, in at least one case [23], the team has been a long-time participant in MUC and used MUC-style task definitions for the biology tasks, and the authors themselves compare their results across domains.

Assuming performance for biology is lower than performance for news, this raises two questions: first, *why* is the performance lower for biology, and second, what can be done to increase performance? This section outlines a number of differences in the two domains that are known to affect performance. These include relative novelty of the task, imprecise task definition, and insufficient training data.

3.1. The experience factor

We know that results for named entity tagging in new languages (Japanese and Chinese, for example) seem to take several evaluations to catch up to the results for English; this probably represents time for the development of resources, including components to handle tokenization or segmentation, morphology and lexical resources. Extraction for news has had the benefit of 15 years of experience; on the biology side, shared resources are just emerging. The GENIA corpus [25] is now providing part-of-speech tagged data as well as biological named entity tags for MEDLINE abstracts.¹³ This cor-

pus will support the development of resources tailored to the domain of biology, such as part-of-speech tagging and morphology.

3.2. Training data

There is clearly less annotated training data available for biology than for news. We see from Table 1 that the amount of annotated training data used for biology named entity tagging is an order of magnitude less than what was used for a high performing named entity system for news (30,000 vs. 650,000). Bikel et al. [18] provide insight into the effect of training data on their Nymble HMM system for news. They report figures using variable amounts of annotated training data, ranging from 60,000 words of data to 650,000 words. The performance varied linearly with the log of the size of the training data set, ranging from almost 92% F -measure to 95%. We can compare this to 20,000 words of training data used in [22], which has a performance of 72% F -measure. Furthermore, we also note that the performance of Nymble [18] on Spanish lags behind English at comparable amounts of data. And at 30,000 words, Spanish name tagging is around 88% F -measure. The discrepancy between Nymble English vs. Nymble Spanish may reflect the “experience” factor mentioned above.

For biology, we have the affects of both the experience factor and very limited training data, plus a third factor, novel names. For example, Krauthammer et al. [21] report that the BLAST algorithm identified only 4.4% of the new names correctly (18 out of 409 names not included in the BLAST database). By contrast, Tanabe and Wilbur [26] used a very general approach for identifying gene-or-protein names in text, drawing on a wide range of lexical and ontological resources, with particular attention to identifying names from their occurrence in indicative contexts. Their performance results range from around 66% to 90%, depending on the cut-off score used. This kind of approach would presumably have much broader coverage (recall) but at the expense of some precision. Since new entities are constantly being discovered and named in biology, any high performing system will have to have a mechanism to recognize novel names from context, without being able to look them up in an existing lexical resource. At this point, we can only speculate on whether the “novel name” task is harder for biology or news; it is certainly a factor for both.

3.3. Interannotator agreement and task definition

Interannotator agreement is far lower for the biological tasks than for MUC newswire (F -measure of 84–89% vs. 97% for news—see Table 1). This may be due to the fact that biologists are being asked to perform a linguistic task that is, from their point of view, somewhat

¹² The evaluation was performed on a single article, so there may be considerable variance in the performance of the system on a larger set of articles. This underscores a major issue for evaluation: detailed evaluation of certain tasks requires a fully annotated corpus, which is costly.

¹³ The GENIA corpus version 3.0 is now available at <http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/3.0/GENIACorpus3.0.intro.html>, with annotations for 2000 MEDLINE abstracts.

artificial. Biologists may not need to look at every occurrence of a term in an article. There can be hundreds of mentions of a gene in a full text article. Table 3 lists frequency of occurrence and distribution for some gene names in FlyBase; the gene name “eye-PKC” occurs 109 times in a single article. Furthermore, mentions in some sections of an article (e.g., Results) are much more important than mentions in others (e.g., in the Background section). It may be most useful for the biologist to know simply that a gene is mentioned in an article, or to know that there is experimental data associated with that gene somewhere in the article. These are the kinds of tasks currently done by database curators. For example, every article curated by FlyBase has an associated list of genes that are discussed in the article. This means that biological databases constitute significant resources that have already been assembled and hand-annotated (curated) by experts, e.g., biologists. The trick is to develop a strategy to use these resources for training and testing of information extraction technology.

The work by Craven and Kumlien [27] explores the use of curated biological databases as a source of “ground truth” or gold standard data. Entries in biological databases encode relations using standardized terminology (or controlled vocabularies) and in many databases, each entry is associated with at least one pointer back to a citation of the article (and its on-line abstract) which gave rise to the database entry. By linking the curated entries to the associated abstracts and defining an appropriate extraction task, Craven and Kumlien show that it is possible to make use of “found” annotated training data. We return to this in the next section, where we outline a source of data for the template element task, and discuss an approach to creating large volumes of annotated training data.

Returning to the issue of interannotator agreement and task definition, lower interannotator agreement for the biology data (84–89%) may simply reflect the fact that biologists do not normally do full named entity tagging for articles. This means that there are no sources of “naturally occurring” fully annotated data¹⁴ for training and test sets when building biology named entity tagging systems. However, if we focus on a biologically motivated task with its associated data resources, we should be able to leverage these resources to demonstrate improved interannotator agreement while simultaneously providing large amounts of annotated data.¹⁵

¹⁴ Here “fully annotated” means that every occurrence of a word or phrase in the text is annotated with its semantic class tag. This is typically done using a text mark-up language such as XML.

¹⁵ This philosophy was behind the choice of data sets for the KDD Challenge Cup Evaluation Task 1, chaired by Alexander Yeh; see <http://www.biostat.wisc.edu/~craven/kddcup/tasks.html> for more details.

3.4. A systematic comparison of biology and news

Leaving aside issues of experience, task definition, and training, we now ask how biology differs from news for named entity tagging. Nobata et al. [28] made a systematic comparison of these two domains by constructing named entity taggers for both domains and doing detailed comparisons. For this, they used two small (25,000–30,000 words) corpora (60 news stories words and 100 MEDLINE abstracts). These corpora contained roughly comparable numbers of named entities (3300 for biology and 2200 for news). They implemented both a decision tree system and an HMM. The *F*-measures reported for the HMM system are 78.6% for news and 75.0% for biology; the decision tree system gets lower scores for both. The paper defines and analyzes the information gain with respect to feature sets and shows that the features used in these systems have higher predictive power for the news data than for the biology data. The features used to model the character types (e.g., case, digits, Greek letters)—originally designed for news—work better for news, and do not work as well (show lower information gain) for biology. The same is true for part-of-speech tags: again, these were developed for general English and do not provide as much information for biology. Information gain also suggests that lexical knowledge is the most useful, but also the most dependent on domain-specific resources.

4. Naming biological entities

There are clearly significant differences between biological names and proper names that occur in the news. This section examines issues in biological name formation and the resulting affects on morphology, ambiguity, and synonymy. We then describe a simple pattern-matching experiment in which we make use of existing biological resources (a comprehensive gene list with synonyms, and annotated data), to explore several dimensions of the named entity and template element extraction problems for biology.

4.1. Biological name formation

There is clearly a major difference between biological names (particularly gene and protein names) compared to the person, organization, and location names in the MUC task definition. Gene names are not, strictly speaking, proper names; furthermore, gene names are productive¹⁶ and they are often descriptive. As new

¹⁶ That is, new gene names can be created by following implicit or explicit rules of gene naming; for example, genes in a family can be differentiated by adding Greek letters as prefixes, as in α -catenin and β -catenin.

genes or proteins are discovered, new names are made up for them. For example, the Mouse Genome Database¹⁷ publishes a weekly nomenclature report, which lists the “nomenclature events” for the week, usually around 50–100 events, including new names and names that have been withdrawn. The rules of name formation in biology give rise to a specialized morphology for prefixes and suffixes. Biological names are also routinely contracted into more manageable abbreviations.

Gene and protein names have distinct characteristics that need specialized handling. The model organism databases publish rules or guidelines for gene names. For example, the first two paragraphs for the naming conventions for *Saccharomyces* read as follows:¹⁸

1. The gene name should consist of three letters (the gene symbol) followed by an integer (e.g. ADE12). Dominant alleles of the gene (most often wild-type) are denoted by all uppercase letters, while recessive alleles are denoted by all lowercase letters.
2. The 3-letter gene symbol should stand for a description of a phenotype, gene product or gene function. In addition, we strongly prefer that a given gene symbol have only one associated description, i.e., all genes which use a given 3-letter symbol should have a related phenotype, gene product or gene function.

The peculiarities of gene naming need to be reflected in biology-specific rules for tokenization.¹⁹ Many separators have multiple meanings. For example, slash, brackets and parentheses can all be parts of words or separators between words. It is clear that a high-performing system for biology would benefit from a specially trained set of tokenization rules.

Complex naming and abbreviation conventions (which can differ from organism to organism) give rise to a specialized morphology for biology. For example, the species for a gene is sometimes encoded as a single letter prefix, as in dPHM—for *Drosophila* PHM gene. Suffixes also carry semantic category information—for example, enzymes often end in “-ase.” Ideally, morphology should inform part-of-speech tagging. For example, a term such as “dPhm” might be recognized as a gene name simply on the basis of its morphology; similarly, “caspase” should be recognized as a term for an enzyme. The proper integration of part-of-speech tagging, named entity tagging and morphology becomes particularly important (and tricky) for abbreviations. These often coincide with English words; for example,

“can,” “for,” and “not” are all *Drosophila* gene symbols. Part-of-speech tagging can be difficult for such words, which already occur in the lexicon but with general English parts of speech. However, these words need to be recognized in context as nouns and specifically as gene names.

This brings us to the issue of ambiguity—a major problem for biological names. There are two kinds of ambiguity: the first is systematic ambiguity, such as that found between genes and their associated proteins [29]. In addition, there is the ambiguity between gene names and words in general English. This becomes particularly important for systems that use large lexical resources. As noted in Stevenson and Gaizauskas [30], a larger lexicon may, in fact, degrade performance because it introduces ambiguity between sublanguage term and the general language. If a term is not recognized because it is ambiguous between a general word in English (e.g., “not”) and a sublanguage term (a gene name), then this lowers recall. If occurrences of an ambiguous term are incorrectly recognized as gene names, this will lower precision. We see a dramatic illustration of this in the experiment described below.

An abbreviation normally accompanies any new name in biology; indeed, for gene names, a “short form” is mandatory. Abbreviations and short forms of long names are a major source of synonymy in biology. Abbreviations also exacerbate the ambiguity problem, since abbreviations tend to be short words and coincide with English words, such as “dot” or “asp.” In addition, abbreviations are intrinsically degenerate forms, so that “asp” may have a number of meanings, depending on the particular domain. A search for the term “asp” in Acromed²⁰ returned over 40 possible meanings, including “abnormal spindle protein,” “antisense promoter,” and “ankylosing spondylitis.”

Finally, synonymy is a major issue when using names as indices into the data. For effective retrieval, it is necessary either to map synonyms into a single canonical form, or to include synonyms in a search. This is particularly challenging because names are constantly changing—not only are new names being added, but names are merging and splitting as new information is discovered. What were originally thought to be products from two separate genes may turn out to be the product of a single gene, with two different functions. And conversely, a single gene product may need to be split into a whole family of products, as happens with alternative splicing. Failure to include synonyms leads to recall errors for named entity tagging. For document level tasks (such as template element), failure to recognize synonymy can also lead to precision errors, by

¹⁷ See <ftp://ftp.informatics.jax.org/pub/informatics/reports/index.html#statistics>.

¹⁸ See http://genome-www.stanford.edu/Saccharomyces/gene_guidelines.html.

¹⁹ Tokenization refers to the separation of a sequence of characters into “tokens” or words. Generally, white space is the basic token separator. However, there are many complications. For example, we see both *13.2-kb* and *300 kb*; in this case, we may want to treat the “-” as a word separator.

²⁰ See <http://gungadin.cs.brandeis.edu/~weiluo/main3.htm>.

failing to collapse two alternate mentions of the same entity.

4.2. A lexical-based pattern matching experiment

It would be useful to quantify the effect of these different issues on the difficulty of name extraction in biology, but because they are interrelated, this is an extremely complex task. To explore these issues empirically, we created a baseline named entity extraction system, drawing on the resources we had available through our work on biological databases (specifically, FlyBase). We implemented a simple longest-first lexical pattern matching strategy for tagging gene names using the FlyBase list of gene symbols and synonyms.

We adopted a task done by the FlyBase curators: listing the set of genes associated with each article. This allowed us to create an evaluation that resembles the MUC-style template element evaluation: extracting the key “cast of characters” (genes) in the article and providing a normalized representation for each gene. The FlyBase record of annotated genes serves as a gold standard, against which the automatically generated list of genes can be evaluated.

To generate the output for evaluation, the system recorded each gene name match, along with the canonical gene name associated with that pattern. These matches were then collapsed and duplicates removed, to produce a list of genes in the paper. This output was scored against the “gold standard” curated set of genes for each paper, using traditional recall and precision metrics. Note that the pattern-matching system was performing a named entity tagging task that was then aggregated into the document level results for evaluation at the template element level.

For this experiment, we drew upon a corpus of 862 full text articles that had been used in a previous data mining task.²¹ Each article had an associated list of FlyBase genes for which there was significant information in the paper. As our lexical resource, we made use of the Flybase gene name list and associated synonyms. This list included (as of February 2002) 35,970 gene symbols, and 48,434 alternative names derived from the curated articles.

We had to decide how to handle ambiguous terms in the baseline system, that is, gene synonyms that could be associated with more than one gene. For example, “Clk” is the symbol for one gene (“Clock”) and a synonym for another gene (“period,” symbol “per”). We did two sets of runs; in one set, we hypothesized all categories associated with an ambiguous gene. This would have the effect of increasing recall at the expense of precision, since at most, one of these guesses would be

correct, and the rest would be wrong. The second approach was to make no guess for ambiguous genes. This would reduce recall, but increase precision.

We analyzed the results under a number of different experimental conditions on a small corpus of 86 articles drawn at random from the larger corpus of 862 full text articles for the KDD Challenge Cup.

The results for pattern matching with and without ambiguous terms are shown in Table 2. First, we see that using this dictionary based approach, we failed to find a number of genes listed in the FlyBase. Recall for full text including ambiguous terms was 84% on our small 86 document set; that is, there were about 16% of genes listed by the curators that pattern matching failed to find. This was surprising, since the system did pattern matching using the complete synonym list of gene symbols and names created by the FlyBase curators. If a gene was listed as occurring in an article, its symbol or a synonym should have appeared in the text. For abstracts, the recall was much lower (31%), which was not surprising but interesting in that only about 1/3 of the genes mentioned in the article appear in the abstract. Even more striking was the very low precision: 2% for full text articles and slightly higher (7%) for abstracts including ambiguous terms.

To determine the source of the recall error, we did a small follow-up investigation for 25 randomly selected missing genes from the original collection of 862 articles. It turned out that in 22 out of the 25 cases, the genes were mentioned only in tables that were embedded as separate files in the original document. As a result, they were not downloaded in the initial download of the documents. While these appeared as recall errors in our template element evaluation, these were terms that were simply not present in the text as downloaded. That is, these missing gene names caused recall errors for the template element task, but did not constitute recall errors for the named entity task on the specific texts that we downloaded.

We also observed other sources of recall error related to the downloading of texts. Typography is a major problem when articles are downloaded. For the KDD data challenge cup experiment, we downloaded only articles freely available on the Web in an HTML version. Because FlyBase uses ASCII for its resources, such as the gene synonym list, we converted the HTML version of the full text into ASCII, following the Fly-

Table 2
Recall and precision for full texts and abstracts

		Recall	Precision
Full text	All words	0.84	0.02
	No ambiguous words	0.77	0.05
Abstract	All words	0.31	0.07
	No ambiguous words	0.28	0.17

²¹ See <http://www.biostat.wisc.edu/~craven/kddcup/tasks.html> for more information.

Base typographic conventions. These included ASCII representations for boldface, italics, and super- and subscripts, as well as Greek letters (indicated as “&agr” for “alpha,” etc.). However, these conversions are somewhat tricky and can lead to pattern matching failures. In one instance, the Flybase synonym list included “Diacyl glycerol kinase ‘&egr,’” but the form that appeared in the paper was “diacyl glycerold kinase {epsilon}.” The pattern matching also missed matching “the soluble guanylyl cyclase [alpha.gif] and [beta.gif] subunits”; the closest FlyBase synonym was entered as “Soluble Guanylyl Cyclase ‘&agr’.”

Spelling variants are captured quite well in FlyBase. For example, “FAS III,” “fas III,” and “fas-III” are all listed (see Table 3) as synonyms. However, for multiword names, there are many combinations, and some are missed, e.g., the case variation in “Guanylyl” vs. “guanylyl” in the preceding example.

Novel names should be a source of recall error in most applications. However, this was *not* the case with the particular data set we were using, since the articles had been curated, which meant that all new names in each article had already been added to the FlyBase synonym list.

Finally, in a few cases, we noticed that gene names occurred embedded in images included in the article. From the point of view of automated retrieval, these would require not only downloading the image file, but doing optical character recognition on text embedded in the image—a difficult task.

This analysis enabled us to account for the recall error. The larger portion of the recall error term was due to the fact that the text received by the system simply did not contain the terms. This is certainly a recall error from the curator’s point of view or from the template element perspective. However, from a named entity task point of view, there would be no text instances of the

name to find in these cases. Thus for the purposes of named entity tagging, we estimate that the lexical matching was finding over 95% of the gene names present in the text. The remaining error term was due to pattern-matching failure related to typography, spelling variants, tokenization, and certain kinds of conjoined constructions.

We next analyze the precision results, which were astonishingly low: 2% for full text articles with ambiguous terms, and 7% for abstracts. The degree of ambiguity surprised us, in that we were using the FlyBase synonym list, which is restricted to genes relevant for *Drosophila*. The precision errors came from three main sources. Some error was introduced by our handling of ambiguous terms, where the system guessed all possible meanings. To determine how much this contributed to the results, we did a second set of runs where we made no guesses for any term that was ambiguous. Table 2 has a separate line for this run and we see that recall drops, as expected, from 84% down to 77% for full articles, while precision increases, from 2% to 5%. For abstracts, recall dropped from 31% to 28%, but precision rose significantly: from 7% to 17%. This indicates that ambiguous terms are one source of the precision error, but clearly not the major source.

One known source of precision errors is the gene/protein ambiguity. Since genes are often named after the protein they code for, there is a systematic ambiguity of names. Hatzivassiloglou et al. [29] developed a classifier that was able to approach human performance on this task. Human annotators agreed approximately 78% of the time while the classifier performed at 75% accuracy, using the human tagged data as a gold standard. In our baseline experiment, some of longer incorrect terms (“cAMP-dependent protein kinase”) appear to be examples of terms used as protein names, not gene names.

Table 3
Three-character and seven-character gene names, sorted by percent curated and document frequency

Gene name	% Curated	# Papers	# Mentions	Gene name	% Curated	# Papers	# Mentions
act	0%	55	143	spliced	0%	11	67
Res	0%	60	123	midline	0%	19	270
did	0%	62	209	blocked	0%	22	52
dot	0%	69	392	missing	0%	29	87
Vol	0%	73	106	smaller	0%	29	58
can	0%	85	680	limited	0%	34	44
for	0%	86	5784	reduced	0%	55	250
not	0%	86	2393	similar	0%	84	461
:				:			
yan	50%	6	17	rutabaga	100%	2	19
DER	50%	8	26	eye-PKC	100%	1	109
EGF	43%	14	131	eyeless	100%	1	9
tra	40%	5	67	Fas III	100%	1	1
egl	33%	3	11	fas III	100%	1	3
eya	33%	3	218	fas-III	100%	1	1
JNK	33%	3	49	GM04742	100%	1	6
MAD	33%	3	4	His2AvD	100%	1	1

The largest source of error is due to the false alarms generated by terms that occur frequently in general English but also are listed as gene symbols in the synonym list (“an,” “by,” “for,” etc.). Table 3 illustrates the fact that many gene names are ambiguous and have meanings as normal English words; in some cases, they would be likely to appear in technical material on *Drosophila* (e.g., “spliced” or “midline”). In other cases, they are out-of-domain English words (“rutabaga,” “18-wheeler”). This phenomenon is discussed in experiments described in Proux et al. [20]; they note that 5.6% of the gene names in their experiment appeared in a general English dictionary. To handle these terms, they found it useful to distinguish between words that have a different part of speech in biology vs. general English (e.g., “if”), and words that would not normally appear in biological texts (“ogre”).

Table 3 also illustrates the highly uneven distribution of words listed in the FlyBase gene synonym list. For the selected three-character names, we see some that occur in almost every article in our 86 article sample (“can,” “for,” and “not”). None of these words is curated as a gene name in these papers. By contrast, we see that “eye-PKC” occurs in one paper 109 times, and it is curated as a gene name; we also see gene names that are curated but occur only a few times in a paper: “eyeless” occurs nine times in one paper; “MAD” occurs in three papers (4 mentions total) and is curated in one of the three papers as a gene name.

We verified the effect of these ambiguous words on precision by eliminating certain common words. Table 4 shows the affect of eliminating words of three or fewer characters. We see that in full text, precision increases from 2% to 6%, while recall drops from 84% to 81%. For the abstracts, the changes are more dramatic: precision jumps from 7% to 29%, while recall falls from 31% to 26%.

This baseline experiment leads us to the conclusion that simple longest-first pattern matching performs poorly at template element and gene name extraction. The recall error is largely due to information not included in the downloaded text. However, the ambiguity problems lead to very poor precision. These results are consistent with the findings in Stevenson and Gaizauskas [30], who report on the effects lexicon size for named entity performance. A larger lexicon can introduce am-

biguity and lower overall performance, as we observed here.

There is an interesting corollary to this experiment, which may lead to automatic creation of large scale corpora for named entity tagging. We know that the articles (and particularly the abstracts) have relatively low recall error for *the named entity task*. The list of curated genes could provide a filter that would “license” only genes and synonyms known to occur in the document.²² We hypothesize that this should eliminate the vast majority of false alarms. The remaining false alarms could be further reduced by automated or manual checking of tagged terms known to be ambiguous. This combination of lexical resources, pattern-matching and filtering should make it possible to automatically create large named-entity tagged training corpora with high recall (over 95%) and low false alarm rate.

5. Lessons learned

Names are critical for biology because they provide indices into the results and the literature. Biology needs information extraction technology to help manage the proliferation of names and their synonyms. However, results for named entity tagging in biology have lagged behind those reported for newswire. While it is difficult to accurately assess the state of the field because of lack of any standardized test and training sets, we have discussed a number of factors that could contribute to these differences. First, biology name extraction is a newer task that has not had the benefit of years of research and the creation of resources and infrastructure to support it. Second, there is relatively little fully annotated training data for named entity extraction. This is related to the fact that full name annotation in running text is not a task that biologists do. This is reflected in poor interannotator agreement and difficulty in collecting data. However, by focusing on a set of biologically motivated tasks, such those performed during database curation, it may be possible to use collections of curated data to create large corpora of named entity tagged data automatically.

We then come back to the original question: why is named entity tagging apparently harder for biology than for news stories? Several things are clear. First, the process of biological name formation is quite different than that for person or organization names. Genes and

Table 4
Effect of eliminating words with 3 or fewer character on precision and recall

		Recall	Precision
Full Text	All words	0.84	0.02
	Only words > 3 chars	0.81	0.06
Abstract	All words	0.31	0.07
	Only words > 3 chars	0.26	0.29

²² It is not always the case that all genes mentioned in a paper are listed. The curators list only those genes for which there is some experimental or novel evidence presented. In practice, this covers most of the genes in the paper. However, if the curated list is incomplete, this could lead to apparent false alarms that are in fact gene names. We plan to investigate how often this occurs.

proteins are often named for their functions. The long full names are almost always abbreviated (creating synonyms). And the abbreviations often introduce ambiguity because of overlap with other abbreviations or general English vocabulary. The nomenclature and the abbreviations also introduce differences in morphology, tokenization and part-of-speech tagging that are specific to biology.

Once we have created sufficient resources to model these differences for biology, it seems reasonable to expect that the performance for biological name extraction will achieve levels of performance comparable to that of extraction tasks for newswire. By taking advantage of both biological tasks and biological data sources, we should be able to provide large training and test sets, as well as large amounts of annotated data to improve performance in this critical area.

References

- [1] Blaschke C, Hirschman L, Valencia A. Information extraction in molecular biology. *Briefings in Bioinformatics* 2002;3:154–65.
- [2] Goble CA, Stevens R, Ng G, Bechhofer S, Paton NW, Baker PG, Peim M, Brass A. Transparent access to multiple bioinformatics information sources. *IBM Syst J* 2001;40:532–51.
- [3] Hahn U, Romacker M, Schulz S. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pacific Symp Biocomputing* 2002;7:338–49.
- [4] Raychaudhuri S, Chang JT, Sutphin P, Altman RB. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res* 2002;12:203–14.
- [5] Chang J, Raychaudhuri S, Altman RB. Including biological literature improves homology search. *Pacific Symp Biocomputing* 2001;6:374–83.
- [6] Masys D. Linking microarray data to the literature. *Nat Genet* 2001;28:9–10.
- [7] Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein–protein interactions. *Int Conf Intell Syst Mol Biol* 1999:60–7.
- [8] Krauthammer M, Kra P, Iossifov I, Gomez SM, Hripsak G, Hatzivassiloglou V, Friedman C, Rzhetsky A. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics* 2002;18:S249–57.
- [9] Hirschman L, Gaizauskas R. Natural language question answering: the view from here. *Nat Language Eng* 2001;7(4):275–300.
- [10] Hirschman L. The evolution of evaluation: lessons from the message understanding conferences. *Comput Speech and Language* 1998;12:281–305.
- [11] Sundheim B. Overview of the results of the MUC-6 evaluation. In: *Proceedings of the Sixth Message Understanding Conference*. Los Altos, CA: Morgan Kaufman; 1995. p. 13–31.
- [12] MUC-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Defense Advanced Research Projects Agency, 1998. Available at <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html>.
- [13] Chinchor N, Marsh E. *Message Understanding Conference Proceedings: MUC-7*, 1998. Available at <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html>.
- [14] Mikheev A, Grover C, Moens M. Description of the LTG System used for MUC-7, 1998. Available at <http://www.itl.nist.gov/iaui/894.02/related_proceedings/lgt_muc7.pdf>.
- [15] Fukumoto J, Masui F, Shimcheta M, Saski M. Description of the Oki System as used for MUC-7, 1998. Available at <http://www.itl.nist.gov/iaui/894.02/related_proceedings/oki_muc7.pdf>.
- [16] Yu S, Bai S, Wu P. Description of the Kent Ridge Digital Labs System used for MUC-7, 1998. Available on-line at <http://www.itl.nist.gov/iaui/894.02/related_toc.html>.
- [17] C. Aone, L. Halverson, T. Hampton, M. Ramos-Santacruz, SRA: description of the IE2_System used for MUC-7, 1998. Available on-line at <http://www.itl.nist.gov/iaui/894.02/related_toc.html>.
- [18] Bikel D, Schwartz R, Weischedel R. An algorithm that learns what's in a name. *Machine Learning, Special Issue on Natural Language Learning* 1999;34:211–31.
- [19] Fukuda K, Tsunoda T, Tamura A, Takagi T. Toward information extraction: identifying protein names from biological papers. *Pacific Symp Biocomputing* 1998;3:705–16.
- [20] Proux D, Rechenmann F, Julliard L, Pillet V, Jacq B. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. In: *Proceedings of the 9th Workshop on Genome Informatics*; 1998. p. 72–80.
- [21] Krauthammer M, Rzhetsky A, Morosov P, Friedman C. Using BLAST for identifying gene and protein names in journal articles. *Gene* 2000;259:245–52.
- [22] Collier N, Nobata C, Tsujii J. Extracting the names of genes and gene products with a Hidden Markov model. In: *Proceedings of COLING '2000*; 2000. p. 201–7.
- [23] Gaizauskas R, Demetriou G, Artymiuk PJ, Willett P. Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics* 2003;19:135–43.
- [24] Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics Suppl* 2001;1:74–82.
- [25] Ohta T, Tateishi Y, Collier N, Nobata C, Tsujii J. Building an annotated corpus from biology research papers. In: *Proceedings of COLING 2000 Workshop on Semantic Annotation and Intelligent Content*; 2000. p. 28–34.
- [26] Tanabe L, Wilbur J. Tagging gene and protein names in biomedical text. *Bioinformatics* 2002;18:1124–32.
- [27] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. In: *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*; 1999. p. 77–86.
- [28] Nobata C, Collier N, Tsujii J. Comparison between tagged corpora for the named entity task. In: *Proceedings of ACL 2000 Workshop on Comparing Corpora*; 2000. p. 20–7.
- [29] Hatzivassiloglou V, Duboue P, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics* 2001;17:S97–S106.
- [30] Stevenson M, Gaizauskas R. Using corpus-derived name lists for named entity recognition. In: *Proceedings of the Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL-2000)*; 2000. p. 290–5.