Annotation Guidelines for Machine Learning-Based Named Entity Recognition in Microbiology

Claire Nédellec, Philippe Bessières, Robert Bossy, Alain Kotoujansky, Alain-Pierre Manine

MIG-INRA, Domaine de Vilvert, F-78352 Jouy-en-Josas
email: forename.name@jouy.inra.fr

Abstract. Recent challenges on machine learning application to named-entity recognition in biology trigger discussions on the manual annotation guidelines for annotating the learning corpora. Some sources of potential inconsistency have been identified by corpus annotators and challenge participants. We go one step further by proposing specific annotation guidelines for biology and evaluating their effect on performances of machine learning methods. We show that a significant improvement can be achieved by this way that is not due to the feature set neither to the ML methods.

Keywords: Named-entity recognition, annotation guidelines, machine learning, biology.

1. Introduction

Named entities (NE) and terms represent the linguistic expressions that denote the objects and concepts in documents. As such their automatic annotation in document collections is a preliminary but crucial step for the semantic annotation and further document processing. Information Retrieval, Information Extraction (IE) and Question/Answering among others, rely on a proper identification of the objects and concepts in the documents. The NE dictionaries and terminologies that are needed for document annotation are available in some specific domains such as biology, but they often suffer from various limitations:

- they are incomplete with respect to the information processing tasks,
- additional disambiguation patterns are needed to handle the ambiguity and polysemy issue,
- variants of canonical terms and named entities that are needed to handle the synonymy issue are missing.

Automatic corpus-based acquisition of new NE and terms, disambiguation patterns, synonyms and variants has been considered as an attractive solution since the beginning of the nineties.

More recently the recognition of biological entities in scientific papers has been popularized by challenges such as NLPBA [Kim *et al.*, 2004 ; *Collier et al.*, 2005] and BioCreative Task1a [Tanabe *et al.*, 2005 ; Yeh *et al.*, 2005]. As for MUC in the news wire domain, publicly available datasets and evaluation reports in biology have a very positive effect on research development in Machine Learning. However, as pointed out by [Tanabe *et al.*, 2005], [Dingare *et al.*, 2004] and [Alex *et al.*, 2006], it is difficult to build a consistent annotation of the training corpus in biology and this negatively affects the reliability of the method evaluation and comparison. Available corpora suffer from various inconsistencies. They are revealed by the analysis of the errors done by the learned NE recognition (NER) patterns when applied on test sets. The sources of potential errors are mainly the fuzzy frontier between entities denoted by proper nouns and entities denoted by terms (compound nouns), the lack of specification of the generality level of the objects to be recognized (entities *vs.* concepts) and the well-known problem of name boundaries. We have thus specified strict guidelines that make the manual annotations easier and more consistent and the NER patterns more learnable. Our strategy consists of splitting the NER task into two separate recognition subtasks, the recognition of the entities themselves and the recognition of their types (*e.g.* *GerE* and *protein* in *GerE protein*). The experiments reported here have been done on the classical problem of the recognition of new gene and protein names in the microbiology domain. We get much better results on the first subtask (*i.e.* entity recognition) than similar methods applied on biology corpora where the distinction between the annotation of entities and types is not so clear.

Section 2 motivates our annotation strategy as derived from the analysis of annotations inconsistencies in available corpora and from previous work on annotation guidelines. Section 3 reports on the experimental results and discuss them with respect to previous results in biology.

2. Annotation guidelines

2.1 NE *versus* terms

The distinction between entities and terms is recent and not fully linguistically relevant but it is operationally useful in IE where NER is one of the main tasks. The acquisition methods differ because of their morphological differences. Named-entities are proper nouns that often have upper case initials. Their variations are mainly typographic (*e.g. sigma K/ sigma(K)*). Terms are common nouns, often compound nouns, which follow traditional inflexion rules and their variations are mainly morpho-syntactic. The following four biological terms illustrate this:

ResD protein, either phosphorylated or unphosphorylated /
both unphosphorylated and phosphorylated ResD /
the phosphorylated form of ResD /
ResD~P.

In NER, the morphology usually determines the conditions that a given name should verify to be recognized as a NE rather than a term: NE recognition is mainly based on typographic criteria. Syntactic criteria have few effects on the NER performance. In biology, this usual morphology-based distinction does not apply. Terms often include proper nouns (Figure 1). Their role is generally to specialize the term meaning by

denoting specific identifiers as in *Streptococcus agalactiae NEM318 serotype III* where *NEM318* and *III* denote the reference to a *Streptococcus agalactiae* strain.
Moreover, the morphology-based distinction does not always fit the semantics; NE as proper nouns can denote concepts or types as well as instances of the concepts. Proper nouns and common terms can even be synonymous and then occur in similar contexts in corpus. Sense disambiguation (attaching the correct type to a given name) and new name recognition cannot rely on the morphology only but also on the context analysis in corpus. Therefore, NE (proper nouns) and term recognition patterns share similar contextual conditions. The example of acronyms and abbreviations clearly falls into this category. *glucose-specific enzyme II (EIIGlc)* where *glucose-specific enzyme II* is a term and *EIIGl* is its synonymous acronym is a representative example of this phenomenon. The NE and the term will be both recognized as enzymes. Typographic criteria are then not sufficient in biology for recognizing named entities.

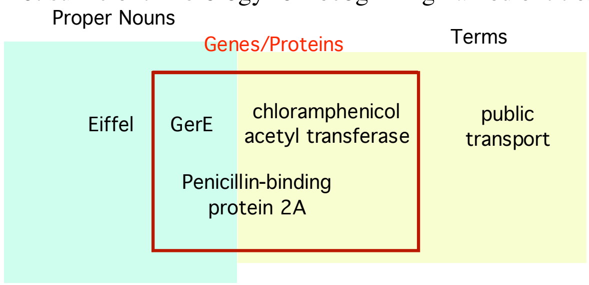
Proper Nouns
Genes/Proteins
Terms
Eiffel
GerE
chloramphenicol acetyl transferase
public transport
Penicillin-binding protein 2A

Fig. 1. Named-entities, Terms or Proper nouns?

In reality, the lexical frontier between the two kinds of knowledge is fuzzy and difficult to formalize. This affects the performances of the Machine Learning methods that are used for learning NER patterns because it is difficult to specify strict annotation guidelines so that the annotation can be reproducible and the NER patterns are learnable.

On the one hand, from the domain expert point of view the guidelines should refer to a consistent semantic category, for instance, *all company names* or *all gene/protein names* independently of their morphology. Such guidelines can make the learning task difficult because the morphologic constraints to be learned are different for the two classes of positive examples, NE and terms.

On the other hand, typographic conditions in recognition patterns are obviously much easier to learn if the guidelines are strict on the morphology - only proper nouns should be annotated as positive examples - but then, the contextual clues required for disambiguating the sense will be more difficult to learn, since terms considered as negative examples share the same contexts as positive examples.

The first strategy has been chosen in previous challenges and evaluations. In BioCreative, for instance, *SAA* and *serum response factor* (respectively proper noun and term) are both annotated as NE proteins. In BioLNLP, *PuB1* and *purine-rich binding sites* are both annotated as DNA (genes). It is natural from an application point of view: what one wants to acquire is a dictionary of a complete semantic category, independently of the morphology. However, the best scores in BioCreative are around 80% recall and precision and 76% recall and 69,4 % precision in NLPBA.

These relatively low scores compared to NE task in MUC can be explained by the morphologic difference of the names to be recognized.

We have thus explored the second strategy, *i.e.* learning proper noun recognition rules. Our hypothesis is that the different types of names, proper nouns and terms should be learned separately from different training corpus and with different methods. The target named entity dictionary would be then built by merging the results of the different learning tasks.

Since some terms include proper nouns, we have specified detailed guidelines so that the annotators can take consistent decisions. Terms that include proper nouns are annotated as named entities, when they denote specific objects and not general categories or types as detailed in the next section.

2.2 Entities versus concepts

The lack of clear distinction between entities (instances) and concepts (types, categories) is another source of inconsistent annotation and machine learning errors. General categories of biological objects are denoted by terms that occur in different contexts than the terms denoting the entities. They are very often in a coreference relation, mainly apposition as pointed out by [Vlachos *et al.,* 2006]. For instance in

> *[...] two alkaline phosphatases (APases) (PhoA and PhoB), an APase-alkaline phosphodiesterase (PhoD), a glycerophosphoryl diester phosphodiesterase (GlpQ), and the lipoproteinYdhF were identified […]*

the entity name *PhoA* is in apposition relation with the concept name *alkaline phosphatase, APase-alkaline phosphodiesterase* with *PhoD* and *glycerophosphoryl diester phosphodiesterase* with *GlpQ*.

Then learning relevant contextual conditions from mixed annotations of concepts and entities at different level of generality is difficult. Moreover, the frontier of the semantic category is much harder to specify in the annotation guidelines, if concepts are included. In biology, concepts are often denoted by general properties as it is in *binding-protein-dependent transport systems* and potentially not useful from an application point of view (*e.g. DNA-binding protein*). The decision to annotate a given term as a relevant concept or not is then difficult to take and very annotator-dependent. What is the limit between entities and concepts in the list *heat-shock sigma factor sigma 32*, *heat-shock sigma factor*, *heat-shock transcription factor*, *stress transcription factor*, *transcription factor*, *factor*? The usual strategies in previous work include both objects and concepts (*e.g. purine-rich binding sites* in NLPBA and *mouse synaptophysin gene* in BioCreative[1]).

We have followed another approach. Only specific objects are considered. For instance, *penicillin-binding protein 2A* is a positive example of protein while *penicillin-binding protein* is not, because it is too general and denotes a *family* of proteins. Following our guidelines, only the first element of the *factor* list above is considered as an entity (*i.e. heat-shock sigma factor sigma 32*). Note that this strategy partly resolves to the problem of the annotation of coordinated noun phrases pointed

[1] The task description mentions explicitly that *human gene* is too general. This illustrates how the limit is hard to specify.

out by [Alex *et al.*, 2006]. In *anhRad54 and hRad54B proteins* we annotate separately *anhRad54* and *hRad54B* and not *proteins*. The problem of annotating intersecting and non-contiguous noun phrases is then overcome. Some coordination problems still remain unsolved as in *interleukin 1 and 2*. Correct annotation of both *interleukin 1* and *interleukin 2* supposed that noncontiguous and intersecting annotations could be made. Note that it is not an inconsistency problem but a problem of specifying an appropriate syntax for the annotation.

Moreover, we have experimentally observed that specific objects (genes, proteins and species) are usually *not* denoted by common terms but either by proper nouns or by *mixed* terms that include proper nouns as identifiers. The morphology distinction looks then consistent with the entity/concept distinction.

2.3 Setting boundaries

The determination of the boundaries is a well-known source of errors. The most prevailing problem in biology is due to the term that denotes the semantic category in the context of the name to be recognized. It can occur before, as a modifier, or after, as a head (*e.g. GerE protein*, *protein GerE*). In most of previous works including NLPBA, the category has been considered as part of the entity name when the name is not an apposition in parentheses, or preceded by a comma. *cAMP regulatory element binding protein* is annotated as a unique name, as well as *a protein kinase A*. The two names in apposition are distinctly annotated in *monoclonal antibodies (mAb)* and in *GATA-1, an erythroid transcription factor*. This results in inconsistent NER where the type of the named entity can belong or not to the recognized name, depending on the punctuation marks of its context.

On the other hand, in BioCreative, the whole noun phrases are annotated even when commas or parentheses indicates chunk boundaries as in *Varicella-zoster virus (VZV) glycoprotein gI* that is annotated as a single named-entity. [Yeh *et al.*, 2005] hypothesized that the lower BioCreative results compared to similar tasks from MUC news wire domain could be explained by longer names in biology. The boundaries would be then more difficult to identify.

To overcome this problem, we follow a different strategy. As stated in the previous section, the expert does not annotate the general terms in apposition relation, such as *monoclonal antibodies* in *monoclonal antibodies (mAb)* but just the entity *mAb*.

Then two cases are considered, either the term denoting the semantic category is the head of the term containing the name, or it is a modifier. In the first case the head is not annotated as part of the entity name. For example, in *cAMP regulatory element binding protein,* only *cAMP* is annotated, as well as in, *Crp*/*Fnr family, the NtrB*/*C two-component system, P78 ABC transporter* (the entity names are in yellow). The short name is considered here as sufficient for naming the object.

In the second case where the semantic category is a modifier as in *cytochrome P450 102* and *penicillin-binding protein 2A*, the semantic category is annotated as part of the name only if it is required for the meaning, as it is the case in the second example but not in the first. *2A* is indeed not sufficient for denoting the protein, while *cytochrome* is redundant. The decision is based on biology expertise: is the category part of the name or not? In fact, the category is usually needed when the name is local to the abstract (as *2A*). Then the name is generally very short and either a simple acronym or mostly composed of digits. Typographic criterion can then help in their identification. To summarize, the name denoting the entity should be annotated without its semantic type except when it is needed for comprehensibility reason. This guideline simplifies the annotation boundary problem and appears as intuitive for most of the biologist annotators in our experiments.

2.4 Semantic type

The last source of error is domain-dependent. The frontier of the semantic category to be annotated is often fuzzy as gene and protein categories are. We have decided to annotate the gene and protein category in their broad sense, including the following objects:

- the objects composed of protein and genes: *loci, alelles, operons*, *gene families*, *regulons*, *clusters, group, regions* and *fusion*
- the subpart of protein and genes: *promoters*, ORFs, *terminators,residues, motifs, boxes,* and *domains*
- part of the experimental material: *reporter genes*, *restriction enzymes*, *restriction sites*, *insertion elements*

A more detailed subtyping is left to further tasks.

The complete guidelines are available at genome.jouy.inra.fr/texte with more examples. The application of these guidelines to a corpus in microbiology is described in section 4. Section 3 presents the machine learning approach and the example representation language.

3. Machine-learning for NER

Our purpose is not to improve ML methods but to measure the effect of the guidelines on the NER performances. In our experiments, we have then selected the most successful approaches as reported in the related work. Previous works differ by the example feature sets, the use of external resources (dictionaries) and the ML method.

3.1 Related work in NER in biology

The main approach in NER in biology until the recent Machine Learning challenges was based on hand-coded pattern design. It relies on multiple sources of information: existing dictionaries and lexica such as UNIPROT, TREMBL, HUGO, UMLS among others [Rindflesh *et al.*, 2000; Cohen *et al.*, 2002; Leonard *et al.*, 2002], character and word-based approaches, linguistic processing [Proux *et al.* 1998], contextual disambiguation and domain knowledge [Humphreys *et al.* 2000; Fukuda *et al.* 1998; Hishiki *et al.* 1998; Franzen, 2002; Narayanaswamy *et al.*, 2003].

Until recently, the ML approach tended to use the linguistic information from the text but only few external resources. It was mainly achieved by the group of the GENIA project [Collier *et al.*, 2000; Nobata *et al.*, 1999; Takeuchi and Collier, 2002; Kazawa *et al.*, 2002]. Recent work agrees on the importance of example representation richness and the central role of the typographic features (see NLPBA and BioCreative conclusions). Among the most relevant features, the case and the non-alphabetic characters (*e.g.* hyphen, digits, symbols) and to a lesser extent, the neighborhood are determinant compared to syntactic categories [Collier and Takeuchi, 2004]. Syntactic dependencies are useful when semantic relations can be derived from them as described in [Wattarujeekrit and Collier, 2005].

Various ML and statistics-based methods have been tested, mainly Markov models, SVM, Maximum Entropy, naïve Bayes and decision tree algorithms. The best scores of the NLPBA challenge [Kim *et al.*, 2004] on the GENIA corpus have been obtained by [Zhou *et al.*, 2004a]. The method reaches 76,0 precision and 69,4 recall. It uses a rich example representation feature set and combines successively HMM and SVM. The best scores of Task1a at BioCreative were obtained by [Zhou *et al.*, 2004b] with a combined approach of HMM and SVM and by [Dingare et al., 2004] with a conditional Markov Model. Both yield around 82-83% recall and precision.

We have designed a similar feature set and selected SVM, C4.5 decision tree method and naïve Bayes (NB) as ML algorithms. We have applied the versions available in the WEKA library with the default parameters.

3.2 Dataset

Our training dataset is a subpart of an initial PubMed corpus on *Bacillus subtilis* (*Bs*) and transcription[2]. *Bacillus subtilis* is a model bacterium that has been extensively studied. The available knowledge on *Bs* genes, functions and metabolism can be usefully exploited for validating information extraction from text. We have chosen this domain because of our deep expertise on microbiology and on this specific *Bs* corpus. Therefore, we have been able to finely control the types of the biological objects to be annotated as well as the level of expert agreement on the annotation. The focus on the transcription issue increases the density of gene and protein names. With respect to the specific issue of transcription, we did not distinguish between genes and proteins as in BioCreative because they often cannot be automatically discriminated by their context because biologists consider the distinction as irrelevant and often use metonymies. A careful analysis did not reveal any obvious complexity difference between the names of our microbiology corpus and those of eukaryote corpora.

431 abstracts have been randomly selected among the 22397 references of the *Bacillus subtilis transcription* corpus. Among them, nine have been manually removed because of their heterogeneity. Their main topic was not microbiology but eukaryotic biology (*e.g. mycobacterium in tumor necrosis mice*). The remaining training corpus then contained 422 abstracts.

3.3 Corpus preparation

For saving manual annotation time, the corpus was first automatically pre-annotated

[2] The query was "Bacillus subtillis AND (transcription OR promoter OR sigma factor)"

by mapping a dictionary of gene and protein names. It was then manually corrected by biologist experts. This strategy is usual in NER. It globally improves the annotation quality but biases the annotation by preferring dictionary names which has a positive effect in our case. We have automatically designed the dictionary in order to limit the number of corrections to be done by the experts. The dictionary contains GenBank gene names of the only species mentioned in the corpus. We have assumed that no gene/protein name would occur in the corpus without a link to its species, except some experimental material such as *lacZ*. This limits the number of potential ambiguities and errors. As such, the dictionary still contained incorrect names because the format and guidelines for entering new references in GenBank are not strictly followed by the contributors. The dictionary was filtered by an anti-dictionary that contained the most frequent ambiguous names, such as *the* and *has* which are actually correct names but also highly ambiguous. It has been completed by six regular expressions that exclude the names represented by one or two letters or digits and long compound terms. The direct mapping of the dictionary to the corpus was completed by typographic variations. The anti-dictionary plus the regular expressions matched 25 014 occurrences in the corpus while the filtered dictionary matches 9 051 occurrences of species and gene/protein names. The number of potentially noisy occurrences was then more than twice the number of the potentially correct ones.

Table 1. Dictionary size.

Number of species names (including variations)	857 451
Number of protein/gene names (including variations)	401 790
Anti-dictionary size	289
Number of names removed by pattern matching	433

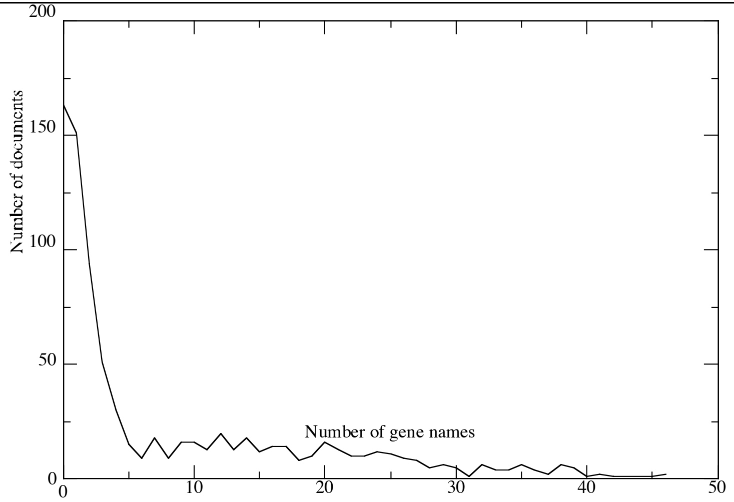

Figure 2. Number of abstracts (Y) containing X *gene/protein names.*

The annotation density varied among the abstracts. Figure 2 shows the distribution of the occurrences of the gene/protein names. Table 2 reports the number of protein/gene names automatically tagged in the corpus.

Table 2. Number of gene/protein names in *Bs transcription* corpus.

# *protein/gene* names occurrences	7 049
# *protein/gene* distinct names	1286

3.4 Manual annotation

The manual annotation was done with the Cadixe XML editor[3]. At the first stage, the corpus has been split into ten disjoint subparts and ten expert biologists corrected the automatic annotation of the dictionary mapping. Then one expert biologist carefully checked the annotation. In case of disagreement, a group of three biologists took a final decision. This way, a full agreement on all annotations was reached. This protocol has been applied for practical reasons only. Independent annotations should be done for measuring the expert agreement.

Three types of corrections of the automatic tagging were performed (Table 3):

- The annotations of the irrelevant homonyms were removed (for instance, *map*)
- The relevant anti-dictionary names (including regular expressions) were annotated (for instance, *has* gene). The length of most of them was one to two characters. The fourth column records those that are more than 2 characters long.
- The relevant names that were not in the dictionary were annotated (referred to as *new names*).
- The boundaries of the names have been modified.

Table 3 reports the number of manual corrections performed by category of error. These numbers are particularly important since they represent the goal of the learning approach: learning rules able to correct as much as possible the annotation done by a direct dictionary mapping.

Table 3. Manual corrections.

	Remove of irrelevant homonyms	Total additions	Addition of anti dictionary names	Addition of anti-dictionary names > 2 char.	Fully new names	Incorrect boundaries
# occ. (1st stage)	1057	1065	123	5	942	714
# occ. (2nd stage)	95	390	177	15	213	154
Total # occ.	956	1276	186	13	1090	781

[3] http://caderige.imag.fr/Cadixe

The number of ambiguities (false positives) was rather high (first column): 13 % (956/7049) of the annotations despite of the use of the anti-dictionary, which has been designed for reducing the ambiguities. The missing annotations were also close to 17 % of the total number of annotations and only a few of them (3 %) were present in the original dictionary and filtered by error. The other errors were due to fully new names, not present in the dictionary. This suggests that the anti-dictionary was not too strict. Incorrect boundaries represented a large part of the errors, around one quarter.

Table 4 reports the final numbers after manual correction. The figures in parentheses represent the name additions compared to automatic annotation. Additions represent the total of the name additions minus the deletions.

Table 4. Manual annotations of the *Bs transcription* corpus.

Total # protein/gene names occurrences	7 185 (+ 137)
Total # protein/gene distinct names	1647 (+ 361)
Total # species names occurrences	2 219 (+ 217)
Total # species distinct names	442 (+139)
Total number of occurrences of NE	9405 (+354)

Table 5 gives the recall and precision measures for the automatic filtered dictionary mapping compared to the manually annotated corpus. The measures were computed as a baseline for further comparison with the ML approach. We counted incorrect boundaries as two errors when an automatic annotation was replaced by one (one false positive, one false negative), three errors when the automatic annotation was replaced by two manual annotations (one false positive, two false negatives) and three errors when two automatic annotations were replaced by one manual annotations (two false positives and one false negative).

Table 5. Precision and recall of the filtered dictionary mapping.

Precision	Recall
76,1	78,1

The performances were surprisingly good compared to previous results by other authors, including the results obtained by hand-coded patterns. The way the dictionary has been filtered by choosing the names related to the relevant species and then filtered by the anti-dictionary was clearly very efficient.

The role of Machine Learning at this point is then double: disambiguating the homonyms and improving the coverage by recognizing new names.

3.5 Example representation

As other authors before, we hypothesized that typographic, linguistic and domain-specific features of the NE and their neighborhood are relevant for designing discriminant NER patterns. Table 6 describes the feature set.

Table 6. Features set

Features

Document structure

- **In_title**: the example belongs to the title.

Typographic features (boolean except length)

- **First_upper**: the example is capitalized (^[A-Z])
- **Middle_upper**: the example contains a non-initial uppercase letter (^.+[A-Z])
- **Only_upper**: all letters of the example are uppercase? (^[A-Z]*$)
- **Last_digit**: the last character of the example is a digit? ([0-9]$)
- **First_dash**: the example starts with an hyphen ('-')? (^-)
- **Middle_dash**: the example contains a non-initial hyphen? (^.+-)
- **Paren**: the example contains a paired set of parentheses? (\(.*\))
- **Space**: the example contains a space character (*ie* is the example is compound? ([ ])
- **Length**: number of characters of the example
- **Between_paren**: the example is enclosed between parentheses without any other word (not a regexp)

Dictionary features (boolean)

- **Eq_dict**: the example is a dictionary entry
- **In_dict**: the example is a strict subword of a dictionary entry
- **Eq_anti**: the example is an anti-dictionary entry?
- **In_anti**: the example is a strict subword of an anti-dictionary entry.

Linguistic features

- **Pos_following_X**: morpho-syntactic category of the Xth word following the example. X ∈ [1 .. 5]. Possible values: J (adjective), N (noun), PP (pronoun), RB (adverb), V (verb), O (other).
- **pos_preceding_X**: morph-syntactic category of the Xth word preceding the example.

Domain specific feature

- **Signal_in_following context:** word X from the signal list belongs to the following context of the example (window [+1 .. +5])
- **Signal_in_preceeding context:** word X from the signal list belongs to the preceding context of the example (window [-1 .. -5])

The role of the signal feature was to represent relevant signal words in the close context of the candidate named-entity. In order to define its value domain from the training corpus, we applied feature selection (based on information gain as implemented in WEKA) to the lemma of the predecessor and successor nouns, adjective and verbs of the positive and negative examples. The negative examples for computing feature selection were all nouns, non positive examples, and followed by a word from the signal list (Table 7), manually built for bootstrapping the process.

Table 7. Bootstrapping signal words acquisition.

activation box dependent enzyme expression fusion gene operon polymerase protease protein regulator regulon replication transcription

The size of the window varies from [-1 .. +1] to [-5 .. +5]. We retained the top 50 words for each window size. The most discriminant words differ depending on the position. For preliminary experiments, we did not want to consider exact position of signal words but an unordered set. In order to select the most popular words among the five lists, we retained the words that belonged to at least 2 lists (*e.g.* it must be top 50 in 2-words window AND top 50 in 3-words window). The lists were then manually filtered by two ways: removing the spurious words such as auxiliary verbs (*be, do, have)* the semantics of which is not clear and removing too specific named entities with the exception of "*lacZ*" and "*Pho*" which are known to be within near context of gene names because they are part of the experiment material. The resulting filtered lists of signal words are given in Tables 8 and 9.

Table 8. List of signal words preceding the NE.

RNAse accumulate bacterial call collision contrary electrophoretic enable enzyme estimate expression genome include intracellular likely phosphorylation probe protein-mediated quantitative relative release respond result role second sequence-selective site-directed summary technique three-dimensional variety

Pho activate activation analysis bind box dependent domain electrophoresis encode enzyme expression factor fusion homologue hybridization inhibit lacZ leader mRNA mutagenesis null phosphatase phosphorylated play polymerase protease protein regulator regulons replication reporter repress require responsible site strain substitution subunit synthetase transcript transcription transcriptional two-component

Table 9. List of signal words following the NE.

Pho activate activation analysis bind box dependent domain electrophoresis encode enzyme expression factor fusion homologue hybridization inhibit lacZ leader mRNA mutagenesis null phosphatase phosphorylated play polymerase protease protein regulator regulons replication reporter repress require responsible site strain substitution subunit synthetase transcript transcription transcriptional two-component

Most of the terms looked relevant as belonging to the candidate named-entity context while some others like *null* or *likely* looked more suspicious.

The positive examples were the examples of NE as tagged in the training corpus. Their description was based on their local context. We have considered fixed size windows within sentences boundaries. The negative examples were automatically derived from the annotated corpus as all noun phrases of one, two or three words in the corpus as analyzed by a basic chunker and non positive examples.

3.6 Experiments

Various combinations of example features were evaluated with the three ML methods, C4.5, SVM and NB. We report here the most significant features namely the typography, the signal words, the syntactic category and the dictionary (Table 10). The first three lines report the results computed with the whole feature set. C4.5 significatively yielded the best results. The most discriminant features of the resulting tree were typographic features (the root was the uppercase initial) and equality of a context word to a dictionary entry or inclusion. The rest of the table reports the results obtained by C4.5. As already pointed out in related work, the most discriminant features seemed to be the typographic ones (- 16 % precision and recall as shown in the last table line). The role of the features related to the dictionary was also important since their deletion yielded 5,5 % lack of precision and 2,1 % lack of recall. The POS tag of the neighbor words of the candidate NE seemed to have no effect on the performances.

Table 10. Experiments with 3 ML algorithms and various feature sets.

	Precision	Recall
C4.5	93,6	93,4
SVM	86,2	89,9
NB	82,8	88,1
C4.5 no signal words	92 (-1,6)	93,3 (-0,1)
C4.5 no dictionary	88,1 (-5,5)	91,5 (-2,1)
C4.5 no POS tag	92,3 (-1,3)	93,9 (+0,5)
C4.5 no typography	77,4 (-16,2)	77,0 (-16,4)

The signal words lack of effect was surprising. Further experiments should be done with different sets of signal words on fixed position, since the lists obtained by the procedure of section 3 generated clearly different sets depending on the distance to the NE. At this stage our conclusion on the design of the feature set is very similar to those of previous works. The typography is very determinant while the POS tags seem to be useless.

Apart from the feature set, we evaluated the effect on the performance of the way the negative examples were generated. As such, the two negative and positive example sets were very unbalanced, the negative set being ten times larger. In order to assess the effect of the negative set size on learning, we trained C4.5 with a subset of randomly selected negative examples, such that this subset was of the same size as the positive set. The results did not improve as opposed to what was expected. It strongly affected the precision (77,6) and increased the recall (98,5). Further experiments should be done on intermediate negative example set sizes in order to evaluate the optimal size according to the corpus redundancy. We did other experiments with various near miss generation strategies that did not yield better results.

4. Discussion and conclusion

As expected, our experiments yielded higher performances than those reported by other authors on a similar NER task and on other corpora. They improve the precision of NLPBA best result by 17,6 % while the recall is 24 % better. Compared to BioCreative, the improvement is more than 10 % precision and recall. The main difference is the domain of the corpora (bacteria *vs.* eukaryotes) and the manual annotation rules. The sets of features are very similar. The ML algorithms are WEKA versions with default parameters and they are less sophisticated than the methods applied by previous challenges winners. We hypothesize that such a performance improvement is mostly due to the respect of consistent and strict annotation guidelines by the biologist annotators. The corpora on bacteria and eukaryotes do not look so different with respect to the NER task that it would explain such different performances. In fact, our results reach similar rates as MUC ones on NER of proper noun such as location and person where the guidelines are comparable to ours: only proper nouns are annotated as NE and not general categories (*e.g.* not *town* in *town of Paris* or not *lake region* in *spring in lake region*). Further experiments with the same feature set and ML algorithm should be done on other corpora in order to confirm it.

We defend here the opinion that different types of knowledge, NER patterns for entities and categories should be separately acquired from corpus. It makes the manual annotation easier and the recognition patterns more learnable. We have demonstrated it here for NER pattern learning in microbiology. We have proposed relevant annotation guidelines with respect to this hypothesis. They are specific to biology and remove most of the inconsistencies observed by previous authors, namely, related to boundaries and granularity.

As specified, the NER learning task does not include more general category learning but only specific entities. We believe that it should be done by a separate learning task with more appropriate techniques that NER pattern learning, including ontology learning (Hearst's patterns and semantic distributional analysis) [Nedellec and Nazarenko, 2005] and term extraction methods that take into account morpho-syntactic variations instead of typographic features. Additionally to these acquisition considerations, it is more relevant from a knowledge modeling point of view to isolate the two tasks so that the two different kinds of knowledge, entities and types are formally represented and linked.

References

1. Kim J.-D, Ohta T. Tsuruoka Y., Tateisi Y. and Collier N. (2004). "Introduction to the Bio-Entity Recognition Task at JNLPBA", Collier et al. (eds), *Proceedings of NLPBA workshop* joint to Coling.
2. Zhou, G., Zhang, J., Su, J., Shen, D., Tan, C., 2004. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 20, 1178–1190.
3. Zhou G., Dan S., Jie Z., Jian S., Heng T. S. and Lim T. C. (2005) "Recognition of Protein/Gene Names from Text using an Ensemble of Classifiers", *BMC Bioinformatics Volume 6, Suppl 1.*
4. Collier N., Nazarenko A., Baud R. and Ruch P. (2005). Recent advances in natural language processing for biomedical applications. *Int J Med Inform.*
5. Yeh A., Morgan A., Colosimo M., Hirschman L. (2005). " BioCreAtIvE Task 1A: gene mention finding evaluation", *BMC Bioinformatics* 2005, 6(Suppl 1).
6. Tanabe L., Xie N., Thom L. H., Matten W., Wilbur W. J. (2005). "GENETAG: a tagged corpus for gene/protein named entity recognition". *BMC Bioinformatics 2005*, 6(Suppl 1).
7. Dingare S., Nissim M., Finkel J., Grover C., and Manning C. (2005) “A System For Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations”. *Comparative and Functional Genomics.*

8. Alex B., Nissim M. and Grover C. (2006). The Impact of Annotation on the Performance of Protein Tagging in Biomedical Text. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, Italy.
9. Vlachos A., Gasperin, C., Lewin I., Yamada C., Briscoe T.,"Bootstrapping the Recognition and Anaphoric Linking of Named Entities in Drosophila Articles", *Pacific Symposium on Biocomputing* 11:100-111, 2006.
10. Rindflesch T. C., Tanabe L., Weinstein J. N., Hunter L. (2000). EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. *Proceedings of PSB'2000,* vol 5:514-525.
11. Cohen K. B., Dolbey A. E., Acquaah-Mensah G. K. and Hunter L. (2002). Contrast and variability in gene names. *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*. pp. 14-20.
12. Leonard J. E., Colombe J. B., Levy J. L. (2002). Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics*, 18:1515-1522.
13. Proux D., Rechenmann F., Julliard L., Pillet V. and Jacq B. (1998). Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Informatics.* 9:72-80.
14. Humphreys K., Demetriou G., Gaizauskas R. (2000). Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. *PSB'2000*, 5:502-513.
15. Fukuda K., Tsunoda T., Tamura A., Takagi T. Toward information extraction : identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on biocomputing (PSB'1998),* 1998.
16. Hishiki T., Collier N., Nobata C., Ohta T., Ogata N., Sekimizu T., Steiner R., Park H. S., Tsujii J. (1998). Developping NLP tools for Genome Informatics: An Information Extraction Perspective. *Genome Informatics*. Universal Academy Press Inc., Tokyo, Japan.
17. Franzen K., Eriksson G., Olsson F., Asker L., Liden P. and Coster J. (2002). Protein names and how to find them. *Int J Med Inf.* 67(1-3): pp 49-61.
18. Narayanaswamy M., Ravikumar K. E., Vi jay-Shanker K. (2003). A Biological Named Entity Recognizer. *Pacific Symposium on Biocomputing* 8.
19. Collier N., Nobata C., Tsujii J. (2000). Extracting the Names of Genes and Gene Products with a Hidden Markov Model. *Proceedings of COLING-2000*, Sarrebrück.
20. Nobata C., Collier N. and Tsujii J. (1999). Automatic Term Identification and Classification in Biology Texts. In *Proceedings of the fifth Natural Language Processing Pacific Rim Symposium (NLPRS).* Beijin, China. pp. 369-374.
21. Takeuchi K. and Collier N. (2002). Use of Support Vector Machines in Extended Named Entity Recognition. *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002),* Taipei, Taiwan, August.
22. Kazawa J., Makino T., Ohta Y. and Tsujii Y. (2002). Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the Workshop of the Natural Language Processing in the Biomedical Domain in ACL '02*, Philadelphia, PA, USA, July.
23. Collier N.and Takeuchi K. (2004). "Comparison of character-level and part of speech features for name recognition in biomedical texts". *Journal of Biomedical Informatics* 37, 423-435.
24. Wattarujeekrit T. and Collier N. (2005), "Exploring Predicate-Argument Relations for Named Entity Recognition in the Molecular Biology Domain", *proceedings of the Eighth International Conference on Discovery Science (DS'05).*
25. Nédellec C. and Nazarenko A., (2005). "Ontology and Information Extraction: A Necessary Symbiosis", *Ontology Learning from Text: Methods, Evaluation and Applications.* Volume 123 Frontiers in Artificial Intelligence and Application, P. Buitelaar, P. Cimiano, B. Magnini (eds.), IOS Press, 2005.