Close Integration of ML and NLP Tools in BioAlvis for Semantic Search in Bacteriology.

Robert Bossy[1], Alain Kotoujansky[1], Sophie Aubin[1], Claire Nedellec[1]

[1]INRA-MIG, Domaine de Vilvert
F-78352 Jouy-en-Josas
{robert.bossy, alain.kotoujansky, sophie.aubin, claire.nedellec }@jouy.inra.fr

Abstract. This paper focuses on the use of corpus-based machine learning (ML) methods for fine-grained semantic annotation of text. The state of the art in semantic annotation in Life Science as in other technical and scientific domains, takes advantage of recent breakthroughs in the development of natural language processing (NLP) platforms. The resources required to run such platforms include named entity dictionaries, terminologies, grammars and ontologies. The demand for domain-specific, comprehensive and low cost resources led to the intensive use of ML methods. The precise specification of the ML task goal and target knowledge, and the adequate normalization of the training corpus representation can notably increase the quality of the acquired knowledge. We argue in this paper that integrated ML-NLP architectures facilitate such specifications. We illustrate our demonstration with four representative NLP tasks that are part of the *BioAlvis* semantic annotation platform. Their impact on the quality of the semantic annotation is qualified through the evaluation of an IR application in Bacteriology.

Keywords: Semantic Annotation, Machine Learning, Ontology Learning, Natural Language Processing.

1 Introduction

Despite the growing number of available structured databases dedicated to biomedicine, a large part of the domain knowledge is only available in documents in natural language. Besides, several services centralize publications in Health and Life Sciences. The main one, Entrez PubMed (NCBI), references over 16 millions of papers [1]. However, at the same time, the size of the bibliographic bases grows exponentially and the scope of the scientific questions crosses the traditional boundaries of biologist expertise fields, making classical Information Retrieval (IR) applications no longer sufficient to target the useful and relevant documents. Advanced techniques involving more semantics have to be applied to textual information processing in the biomedical domain.

Life and Health sciences are recognized as critical knowledge-intensive domains for the Semantic Web [2]. Research efforts towards the Semantic Web aim “at replacing the current *web of links* with a *web of meaning*” [3] producing large-scale methods for automating deep semantic analysis and markup of Web pages in a machine-readable form suitable for information extraction (IE) or information retrieval (IR) applications in the biomedical domain. Semantic analysis methods involve more and more Natural Language Processing (NLP) and powerful representation languages that are reaching a maturity stage, where fine-grained semantic markup of large Web corpora of text in various domains and languages become possible. This is demonstrated for instance by the GATE [4], UIMA [5], and Alvis [6] NLP platforms that make an extensive use of linguistic knowledge. A large part of it is domain-specific and requires costly development and maintenance efforts that can be alleviated by Machine Learning (ML) methods. Corpus-based ML methods yield impressive knowledge acquisition results for a wide variety of NLP tasks such as named entity recognition (NER) [7, 8], POS tagging [9] and concept and relation tagging [10, 11]. However, the cost remains high for (i) the production of the appropriate features for representing the training examples, (ii) the manual annotation of the training examples and (iii) the evaluation of the quality of the ML results. The close integration of ML methods and end-user applications, *e.g.* IE or IR, into semantic annotation platforms gives a useful framework to overcome these limitations. Such efficient platform integration implies the proper characterization of the type and role of the knowledge that is used and produced by each platform component. This formalization step allows to avoid many cases of redundancy and inconsistency of semantic annotations. Translating this into ML concerns means that the learning target concept must be clearly specified according to the overall knowledge model and the design of the example representation should be derived accordingly. Following this principle, we have defined four representative and related learning steps and the NLP process that computes the necessary training corpora. Experimental results with the BioAlvis ML-NLP platform show that the appropriate normalization of the example representation according to the learning task improves ML performance and facilitates further knowledge integration. With the application of the BioAlvis platform to IR of biomedical documents, we measure the quality improvement of the semantic annotation performed with the learned knowledge.

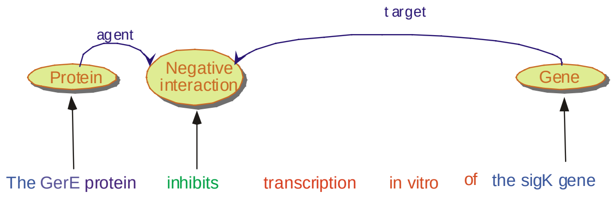
2 From Words to Concepts

2.1 Semantic Annotation

Automatic semantic annotation supplies a meaningful structure to free texts expressed in natural language with the purpose of allowing machine processing. In the Semantic Web framework, the semantic annotation consists in an interpretation of the text supported by an ontology, *i.e.* the assignment of concepts and relations of an ontology to fragments of text. The extent of the annotated text fragments is fairly variable depending on the target application. IE and IR target specific bits of information contained in short fragments of text, *i.e.* terms, words and multi-word units. Fig. 1 shows an example of the semantic annotation of a sentence from a scientific article on Molecular Genetics. The word *GerE* denotes a protein and the word *sigK* denotes a gene. The *negative interaction* concept is supported by the *inhibit* verb. GerE (resp.

sigK) and the verb *inhibit* are instantiations of the arguments of the ontology relation *agent* (resp. *target*) between the protein and the concept *negative interaction*.

Fig. 1. Example of semantic annotation in Biology.



As an illustration of the production and exploitation of a semantic annotation in the context of IR, we present the BioAlvis variant of the Alvis framework focused on bacteriology that performs as follows. The annotation pipeline enriches documents with fine-grained semantic annotations acquired through the successive application of NLP tools. The result is passed to the indexing component and exploited by the semantic search engine. The IR service normalizes the user queries in the same way as the documents: words are lemmatized, terms and named entities (NE) are replaced by their canonical forms and the concepts are replaced by their paths from the ontology root. This strategy differs from usual query expansion that consists in replacing each query term with the set of synonyms and sub-concepts. The Alvis method indeed drastically reduces the complexity queries and makes its interpretation legible for the user. The user can also directly benefit from the annotation of concepts by performing ontology-based facet refinement through a rich Web user interface.

2.2 NLP/ML cooperation towards semantic annotation

Software platforms for text corpus annotation integrate a common range of linguistic processes into pipelines, typically: tokenization, word and sentence segmentation, named entity and term tagging, part-of-speech tagging, syntactic parsing and semantic concept and relation annotation. Each process relies on linguistic resources relevant to the target domain, which requires important acquisition efforts. Most platforms do not specifically include *automatic* knowledge acquisition facilities (*e.g.* Luxid®, MedScan®, AKS2®, InGenuity®) or in a limited way (*e.g.* Luxid I2E® for NER), although corpus-based Machine Learning provides an attractive alternative to the manual acquisition of such resources. Technically, a single annotation pipeline can process documents for application purposes as well as for preparing training corpora with the intent of acquiring new linguistic resources. However, in most implementations, this virtuous feedback does not translate into close ML and NLP software integration. The input of the ML is usually computed by a subpart of the NLP pipeline but the output is not directly usable by subsequent NLP components. This is the case in Gate / Amilcare [4].

We claim that semantic annotation can greatly benefit from a full integration of the ML components that feed the knowledge bases. Beyond format homogeneity, close integration compels the architecture designer to specify the respective roles of each NLP component involved in the semantic annotation process, to identify precisely all types of knowledge along with their interdependencies, and the target knowledge to be acquired by each ML component.

Breaking down the semantic annotation task into well-identified NLP elementary steps has a positive effect on the production quality of the associated ML component. Relevant regularities are more easily identified by the ML system and human annotation of training examples is easier and of higher quality when it concerns a singled out knowledge type. For example, in formal knowledge representation frameworks, the tagging of semantic types and the tagging of properties are two distinct steps. In the phrase, "*mouse* ***synaptophysin*** *gene*", the annotation of ***synaptophysin*** as an object of type *gene* is handled separately from the annotation of its property *belongs to the species mouse*. The knowledge acquisition goals for the recognition of gene names and their related species must be achieved by at least two distinct ML tasks applied to two different training corpora. The increased homogeneity of corpora that results from normalization reduces the number of examples to annotate manually. Unfortunately, many knowledge acquisition approaches to NER do not follow this principle [8].

Moreover, the clarification of dependencies among the different types of knowledge provides a basis for increasing knowledge modularity and reducing annotation inconsistencies. Operationally, the dependencies between knowledge types impose a constrained order of linguistic/semantic/acquisition processing steps that should be made explicit. In a structured modular view of the linguistic knowledge base, higher level knowledge should encapsulate lower level knowledge. Then, in order to learn a given target knowledge *K*, the training example representation should be based on the knowledge on which *K* depends and no any lower level knowledge. For instance, for the sake of modularity, relation recognition rules should not be learned from shallow clues such as punctuation marks. Previous NLP steps should have interpreted the punctuation marks into relevant information such as sentence ends (sentence segmentation) or abbreviations (named-entity normalization).

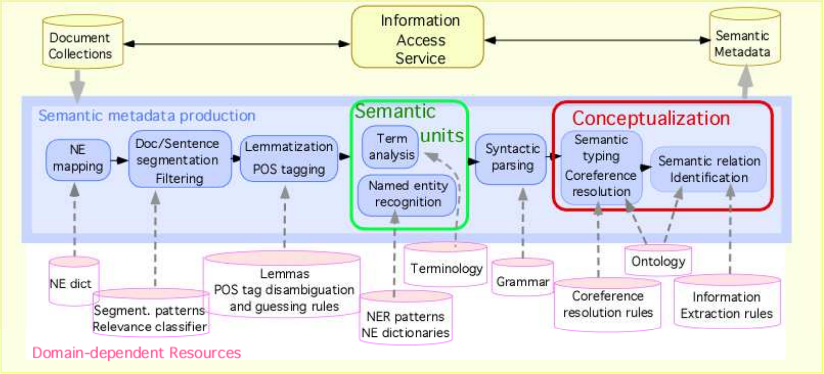
Hereafter, we present the results obtained by applying these principles to the development and the integration of knowledge acquisition facilities into the Alvis platform. We focus on the acquisition of critical resources that are required by semantic annotation, with respect to the variety of learnable linguistic knowledge, *i.e.* named entities, terms, concepts and relations. We demonstrate their learnability (section 3) and the benefit of fine-grained semantic annotation in terms of quality and density of annotations for a given domain and application: IR in Biology (section 4).

3 The BioAlvis Experience

3.1 Architecture

The Alvis annotation/acquisition pipeline (*A3P* in the following) has been developed within the Alvis project [12]. Alvis aimed at developing an open software platform that supports the quick development of distributed semantic search engines in specific domains. Alvis platform integrates a semantic crawler, the annotation/acquisition pipeline and a semantic search engine based on Zebra [13]. As a proof of flexibility, various instances of Alvis have been deployed in a short time for different languages (*e.g.* Chinese, Slovene, English and French) and different genres and domains (*e.g.* textbooks, news, patents, Wikipedia entries, MedLine abstracts, Agrobiotechnology patents). The Biology instance, *BioAlvis* developed by us, is presented here. Following the principles advanced in section 2.2, A3P is composed of a sequence of modules, based on *Ogmios* [14], that produce a layered annotation of the input document (central area of Fig. 2). The modules communicate by the means of a common layered XML annotation format.

Fig. 2. BioAlvis architecture.


Document Collections
Information Access Service
Semantic Metadata
Semantic metadata production
Semantic units
Conceptualization
NE mapping
Doc/Sentence segmentation Filtering
Lemmatization POS tagging
Term analysis
Named entity recognition
Syntactic parsing
Semantic typing Coreference resolution
Semantic relation Identification
NE dict
Segment. patterns Relevance classifier
Lemmas POS tag disambiguation and guessing rules
NER patterns NE dictionaries
Terminology
Grammar
Ontology
Coreference resolution rules
Information Extraction rules
Domain-dependent Resources

The XML annotation format relies on a layered representation where each layer gathers the annotations from a single type of knowledge, in a similar way as described in [15]. The first annotation steps identify *semantic units*, *i.e. named entities* and *terms*, that denote the reference concepts of the domain in the document (*Semantic Units* box of Fig. 2.). Their recognition, their normalization and their disambiguation require prior word and sentence segmentation and word lemmatization. The next annotation steps assign ontology categories to the semantic units (*Conceptualization* box of Fig. 2). This includes fine-grained sense disambiguation based on selectional restriction of the semantic units (section 3.3). Prior syntactic parsing produces the required syntactic dependencies. Finally, ontology relations among the semantic units are identified by the application of Information Extraction rules. The rules use the semantic categories and the syntactic dependency context of the semantic units.

A3P bootstraps by providing an annotated corpus for the acquisition of the knowledge for the next components in the pipeline sequence. As shown in Fig. 2, the linguistic analysis modules are fed by knowledge bases (drums). Their acquisition is achieved by an array of corpus-based acquisition tools involving ML methods.

The next subsections describe four representative acquisition tasks of BioAlvis, their target knowledge, the example representation and the principles of the integration of the learned knowledge into the knowledge bases. Most of the learning results described in the following were obtained from a representative training corpus in bacteriology of 2,397 scientific paper references from MedLine referred to as the *Bacillus* corpus designed in 2001. It is the result of the following query to PubMed: "*Bacillus subtillis AND (transcription OR promoter OR sigma factor)*".

3.2 Semantic Units

3.2.1 Named Entities

The term *named entity* usually designates proper names associated to an ontological category or semantic type (*e.g.* place, person). The proper names are *rigid designators* that denote a referential entity in an unambiguous way [16].

BioAlvis NER component, *RenBio* focuses on protein/gene and species recognition. It achieves NE tagging by GenBank-based dictionary mapping and by the application of disambiguation and variation rules. The disambiguation rules specify what contexts are required for each type of named entity. In parallel, variant dictionaries and variation rules in the form of hand-crafted regular expressions deal with common typographic alterations. Rules for disambiguation and recognition of new entities are automatically acquired by supervised machine learning from a reference training corpus. For example, the simple rule,

A word, followed by the word protein, 4 letters long, starting and ending with an upper-case letter, is a protein name.

applied to the text, "*The* ***GerE*** *protein inhibits transcription*" assigns *GerE* to the *protein* category, even if *GerE* is not in the protein dictionary.

The linguistic features of the training examples are computed by segmentation, lemmatization and typographic analysis (*e.g.* length, case, presence of symbols and digits, co-occurring words) of the training corpus performed by the annotation pipeline. The annotation of positive examples is done by first mapping the NE dictionary and then its manual correction by domain experts. Negative examples are automatically generated under the closed-world assumption.

The *RenBio* rules for gene and protein name annotation in BioAlvis were learnt from the *Bacillus* corpus. The dictionary mapping on the training corpus tagged 7,185 occurrences. 10 biologists analyzed, corrected and completed the tagging. They found 12% false positives due to ambiguities and 12% false negatives due to new names. We applied the C4.5 algorithm of induction of decision trees (J48 WEKA library version [17]) to the revised training corpus. The cross-validation evaluation reported in Table 1 showed significantly better results in terms of recall and precision compared to the best results of the two gene/protein recognition challenges NLPBA [18] and BioCreative II [19].

Table 1. NER performances (recall-precision).

Best *NLPBA*	Best *BioCreative II*	RenBio *Bacillus*
76% - 69%	86% - 88%	94% - 92%

We claim that the example representation features and accurate specifications of the learning goal permitted higher quality of the training examples, thus improving the conditions for the learning algorithms. The good results were not due to any breakthrough in ML since we apply a regular well-known algorithm.

On the one hand, the automatic linguistic pre-processing by BioAlvis of the examples has contributed to drastically reduce the dimension of the example description space and to remove potential sources of errors. Moreover in order to discriminate between NE and non-NE by their context, we picked the most relevant trigger words by feature selection. For instance, words like *gene*, *operon* or *transcription* are more likely present around gene names than any other word.

On the other hand, the learning goal was specified according to the role of the NER in the semantic annotation process. This strongly determines the guidelines for the manual annotation of the training examples by experts. Our strict annotation guidelines address several phenomena that could hinder the quality of the annotation. The principles are as follows: NE annotation should be restricted to single entities for learnability and knowledge modeling reasons, it should exclude terms that denote general semantic categories and properties and the entity span should exclude the description of entity qualifiers (e.g. in *"recA gene"*, only *"recA"* is annotated). The detailed description of the guidelines can be found in [20].

Our experiments demonstrate that the combination of the appropriate normalization of the data with the consistent annotation of training examples by experts improve machine learning performance in terms of precision, recall and size of training sample (see [20] for more details).

3.2.2 Terms

The BioAlvis *term analysis* component identifies the phrases that represent semantic units. These are single or multi-word nominal or verbal expressions that refer to specific domain concepts (*e.g. plant pest*). The term analysis module achieves term recognition and normalization, which consists in tagging the term with its canonical form. Semantic ambiguities are processed afterwards by the semantic typing module (section 3.5), while inflections are handled beforehand by the lemmatization module. Similarly to NE, off-the-shelf lists of terms are not sufficient to annotate documents because terms may be ambiguous and subject to variation. Moreover, in scientific and technical domains, terminologies are generally incomplete with respect to the specific application needs [21]. Thus less than 1% of the 410 000 terms of MeSH [22] and Gene Ontology [23] occur in the 16,000 sentences of the *Bacillus* corpus. In addition to being mostly related to eukaryotes, those terms are suitable for manual indexing as they do not appear as such in NL documents.

BioAlvis includes two acquisition modules for enriching the terminology with new terms and variants from a training corpus. The term acquisition component is the *YaTeA* term extractor [24]. It takes as input a training corpus with segmented sentences and words, lemmas and POS tags. YaTeA identifies candidate terms in an unsupervised way. Its strategy is based on declarative linguistic rules for boundary detection and term analysis and on endogenous disambiguation. Extracted candidate terms are usually validated by domain experts.

Variation spectrum is much larger for terms than for NE. In addition to minor graphical variations, it includes morpho-syntactic alterations that can deeply modify the form of the term (*e.g. plant pest*, *pests on plants* and *pests that attack plants*). Term normalization involves complex linguistic and domain knowledge encoded in variation rules. BioAlvis integrates *FASTR* [25], a tool that computes candidate term variants from training corpora and controlled terminologies as produced by YaTeA and experts. For instance, FASTR insertion rule extracts *genetic competence* from the *Bacillus* corpus, as a variation of *competence*. Domain experts then validate the proposed variation relations as synonymy, hyponymy or other relations. Note that sets of synonym variants of the same term are similar to WordNet synsets. A canonical form is chosen for normalization purposes to represent the synonym set.

Applied to the *Bacillus* corpus, YaTeA extracted 6,699 candidate terms occurring at least twice, among which 3,560 were validated by a group of 3 experts in terminology and biology, yielding 52% precision. The recall evaluated on a gold standard subset of *Bacillus* was 67%. FASTR then extracted 2,335 variants. The validation by two experts was done in a few days and resulted in 676 synonym sets and 1,569 hyponyms. Additionally, from 715 MeSH terms found to occur in the *Bacillus* corpus, FASTR identified 1,899 new variants among which 397 hyponyms and 117 synonyms.

Such methods, when integrated in a pipeline, appear to be very competitive compared to manual acquisition. The approach offers exhaustiveness regarding to a corpus that is a clear advantage for knowledge-based application.

3.3 Semantic Types

Semantic typing relates concepts from the ontology to semantic units in the text after their identification by the term and NE recognition components. The ontology concepts are organized into hierarchies. Semantic typing annotates the semantic units with the whole concept path to the hierarchy root without any a priori assumption on the concept level relevance. In A3P, the ontology-lexicon relation is explicit: the concepts of the ontology are represented at the lexical level by the canonical forms of terms and named entities. In case a given semantic unit is polysemic, disambiguation rules select the right concept in the ontology with respect to its syntactic context in the document; BioLG [26, 27], the dedicated version of Link Grammar, is integrated in BioAlvis for computing the syntactic contexts.

The acquisition of concept hierarchies and disambiguation rules is supported by the ML hierarchical conceptual clustering tool *Asium* [10]. Asium takes as input a training corpus annotated in the same manner as the input of the semantic typing module, *i.e.* semantic unit tagging and parsing. The formation of concept classes by Asium is based on distributional analysis assuming that semantic units occurring in similar syntactic contexts in specific domain corpora are semantically close. Asium suggests their corresponding concepts as candidate members of a same semantic class. The disambiguation rules are automatically learned together with the semantic classes. They are expressed as restrictions of selection stating the syntactic dependency constraints on the context of the semantic unit being typed. For instance, *cat* may both denote a *mammalian* or a *gene* as defined by the ontology. In the phrase *hypokalaemaic myopathy of Burmese* ***cat****, cat* must be tagged as *mammalian* rather than as *gene*. *myopathy* is a *disease* and the disambiguation constraints express the knowledge that mammalians have diseases, but genes have not. Semantic classes are successively merged according to *Asium* distance formula. *Asium* includes a user cooperative interface to validate, revise and name the learned classes and hierarchies on the fly as they are built. This iterative process avoids error amplification along the hierarchy formation. Like all distributional semantics-based methods, *Asium* produces large coverage classes that may include three types of errors: the input syntactic dependencies computed by the parser may be incorrect, (*e.g.* between 20 and 30 % of the dependencies computed by BioLG [27]); the syntactic context may reflect different meanings (*e.g.* the preposition *in* expresses either time or place relation as in

transcription in *mitotic cell cycle* **/** ***transcription in*** *cell*), which implies splitting the class; the semantic relation may not represent only close meanings but antonyms or lexical variations that were missed by the term and variants analysis.

A large part of the potential learning errors is avoided upstream by choosing an appropriate representation of the training examples. Terminology and NE normalization significantly improve the quality of the learned classes by increasing the homogeneity of the training data. It also decreases the number of parsing errors and reduces the computation time, since the parser avoids computing dependencies inside the terms as detailed in [26]. Indeed, normalization removes irrelevant variations by a factor of 3 to 4. Moreover, syntactic contexts as used in *Asium* reflect more accurately the semantic roles than typographic windows can do. Extensive evaluation of the quality of the semantic classes acquired by distributional semantics based methods has not been conducted yet. However, a general comparison of Asium with other systems can be found in [28].

Although the concepts are validated along their construction by Asium, they cannot be integrated as such in the ontology. Their structure does not necessarily represent the model needed for the application and may require validation and revision. The alignment between learned ontology and existing ones also remains a critical problem. The modeling strategy we adopted for the development of *BacteriOntology* is based on an ontology skeleton crafted by hand by biologist experts and computer scientists from MIG-INRA laboratory. The hierarchical model results from the integration of several existing resources: (1) the highest levels of the ontologies and thesauri GO and MeSH; (2) relevant domain-specific information resources (*Riley* and *Subtilist* function classifications and NCBI species taxonomy); (3) concepts denoted by the 300 most frequent terms (section 3.2.2) acquired from our corpus. *Asium* results were then used to extend this core ontology and populate its classes. The current version of the resulting *BacteriOntology* defines 5,888 concepts, structured into 6 generality levels (excluding the extremely deep species hierarchy). The quality of *BacteriOntology* acquired with Asium support was globally evaluated through IR (section 4).

3.4 Domain-specific Relations

Domain specific relations are usually more difficult to identify in the text than concepts because they are less directly supported by contiguous text fragments. BioAlvis annotation of relations focuses on gene interaction and relies on relation extraction rules. For a given relation, the rules check the type of the semantic units in the ontology in order to spot candidate relation arguments, and the type of the syntactic dependencies between them. For instance, in the text: *GerE inhibits the expression of sigK*, the gene interaction between the protein agent *GerE* and the target gene *sigK* is identified in the simplest case, by the rule expressed in first-order logic:

gene_interaction (X,Z) :- type(X,Protein), subject(X,Y), type(Y,Interaction_action), obj(Z,Y), type(Z, Gene).

where *Protein*, *Interaction_Action* and *Gene* are ontology concepts, and *obj* and *subject* are syntactic dependencies. Many complex gene interaction cases are handled with the same method including those involving regulon membership and promoter binding (detailed method in [29]). Relation extraction rules are learned by the supervised Inductive Logic Programming method, LP-Propal. The training examples are expressed in the same way as the input corpus of the relation tagging component: typed semantic units and syntactic dependencies. The sentences are selected by the naïve Bayes classifier STFilter [30], so that manual annotation focuses on the relation arguments in the sentences that most probably express a genic interaction. The successive filtering, disambiguation and normalization of the lexicon and syntactic analysis improve the training set homogeneity.

The LLL dataset on genic interaction [31] has been designed from the *Bacillus* corpus for evaluation. Experiments on the subset *action without coreference* (70 positive examples) yielded 89.4% F-measure. This result is significantly better than the 65.5% best LLL challenge score on the same dataset [32] and than the BioCreative II result (48%) on the protein-protein interaction task [33]. We tested our system with an altered representation of the same data, where syntactic dependencies were replaced by word neighborhood relations. Considering the poor results (34.7% recall and 22.8% precision), we proved that syntactic dependencies convey major semantic relation information.

4 IR Experimentation in Biology

We have designed the BioAlvis version of Alvis for the evaluation of ML-based semantic annotation benefit and the delivery of knowledge-based application to biologists (*e.g.* IR). This section reports on the experimental evaluation of the BioAlvis semantic annotation for semantic search and its comparison to other indexing and search models. We characterize Alvis search as *semantic* in the sense that it automatically interprets the meaning of the query with respect to the terminology and the ontology: Alvis searches for more specific and variant terms and it assists query refinement by ontology and terminology navigation (see [34] for more details). We compare Alvis retrieval capabilities to three representative IR services that are intensively used by specialists in specific domains and particularly in Biology: (1) Google and (2) Google Scholar represent automatic full-text indexing with shallow linguistic processing and (3) PubMed Entrez is representative of hand-crafted indexing by thesaurus keywords and full-text indexing *without* linguistic processing. The comparison focuses on the effect of semantic annotation and query expansion on the answer set quality. We exclude the effects of result sorting (ranking) and of interface facilities (query refinement). Although they are obviously important features, they are outside the scope of the evaluation.

4.1 Test Data

The reported experiments concern the adaptation of *enterobacteria* to changes in their environment. *Enterobacteriaceae* is a large family of bacteria of the intestine, including many human pathogens, like the well-known *Salmonella* that causes inflammation of the intestine (*Gastroenteritis*). Their virulence is due to their capacity to survive and grow in hostile environment conditions imposed by their hosts (acidity, high temperature or oxidative stress induced by iron starvation and superoxide radicals). Part of these conditions is due to the response of the host organisms to pathogen infection. The deep understanding of the bacteria response mechanisms at a molecular level to these stress factors is a key point toward the design of more efficient drugs. The goal of the search is to find descriptions of pathogen reactions and was translated into the following query: *enterobacteria stress genome component*.

In order to test *BioAlvis*, the *Bacteriology* document collection was built by first querying PubMed with all bacterial genera names from the GenBank taxonomy. Then we used a Bayesian filter to exclude documents that were not bacteriology *stricto sensu*. The result is a medium-size corpus containing 322,982 references of 70 words on average. This corpus was processed by BioAlvis in 60 hours on a cluster of 20 processors. The resulting semantic annotation was indexed and supplied to the Alvis search engine. Table 2 summarizes the main figures of the acquired linguistic resources (as described in section 3) and tagging features.

Table 2. Annotation of the Bacteriology corpus.

Type of resource	Size of the resource	Tagging
Gene/protein names	1,686,244 different forms 666,797 canonical forms	2,046,262 occurrences 200,225 different names 12% of the dictionary Avg. 6 gene or prot. names/doc.
Species names	748,262 different forms 270,159 canonical forms	1,309,801 occurrences 30,985 different names 4% of the dictionary Avg. 4 species names / doc.
Terms	7,279 canonical forms	2,449,669 occurrences 5,804 different terms 80% of the dictionary Avg. 7 terms / doc.
Conceptual hierarchy	5,888 concepts (831 > level 0)	2,305,747 occurrences 740 concepts of level > 0 89% of the concepts > 0 Avg. 11 concepts / doc.

The annotation is dense due to the type of documents and the corpus-based strategy of the lexicon acquisition. For instance, the BioCreative II corpus contains on average 4.6 gene or protein names per document while there are 6 in our corpus.

4.2 Compared Systems

Google and Google Scholar index very large collections. Google references around 24 billions of web pages. The size of Google Scholar is estimated at more than one billion references. Both systems perform simple stemming on documents and queries. Our hypothesis is that in specific domains, they will (1) retrieve more incorrect results compared to semantic search, because they do not disambiguate words; (2) miss relevant documents by not exploiting synonymy and related terms.

In Entrez PubMed, each indexed reference is manually assigned a set of terms representative of the document topic from the MeSH thesaurus. The manual annotation avoids ambiguities in document indexing but is quite expensive to maintain since it requires highly-trained experts who read the full-text articles. Entrez PubMed searches query terms in the full-text without any linguistic analysis as well as in the MeSH term index by expanding the query with synonyms and more specific terms according to the MeSH thesaurus. In all cases, the resolution of query ambiguity is postponed to query refinement by the user.

To illustrate the strategy of BioAlvis, we detail how the example query *enterobacteria stress genome component* is processed: the words are lemmatized; the recognized semantic units are normalized and assigned to the *BacteriOntology* concepts that belong to taxonomies: *enterobacteria*, *stress* and *genome component*. BioAlvis expands the search to documents where sub-concepts of the query term occur. For instance, the taxonomic group of *enterobacteria* contains *Escherichia coli* and *Salmonella enterica* among hundreds of other bacteria species. In the same way *stress* defines 17 different types of stress such as *heat-shock* and *phosphate starvation*. Again, *genome component* represents 62 different sub-concepts (*e.g. operon* and *promoter*). Each of these concepts references its variants and synonyms. For instance, *heat-shock* is synonym of *temperature upshift*, *thermal upshock*, and *temperature upshock* according to our terminology. Additionally, query lemmatization allows BioAlvis to search regardless of word inflections and derivations (*stress* / *stressing* / *stresses*). The interface displays the detail of the interpretation so that the result is understandable.

4.3 Experiment and Evaluation

As we could benefit from Bacteriology expert analysis, we opted for a qualitative evaluation of our system. Beyond rough figures, a comparative study of the answer sets has characterized the missing and irrelevant documents retrieved by each service. More complete results can be found in [34]. Table 3 summarizes the features of the answer set for the four IR services, including Alvis.

The very large answer set of Google (245,000) was expected because of the document collection size and the generality of the query. Google and Google Scholar search for the stemmed query words in the documents. As no semantics is used, all documents with sub-concepts of the query words were missed. We tested the query with a replacement of *genome component* by *(gene OR promoter)* that are two productive sub-concepts and found that 50 % more documents were retrieved. As Google and Google Scholar make use of stemming, they find 8 times more documents than with exact matches. For instance, documents with *enterobacterial* were found thanks to stemming.

Table 3. Size of the query answer sets for tested search service.

Google	*Google Scholar*	*Entrez PubMed*	*Entrez PubMed*	*BioAlvis*

			w/o MeSH	
245,000	2,740	1031	0	1,870

Entrez Pubmed expands queries in a similar way as BioAlvis by following MeSH relations top-down. The term *enterobacteria* is expanded into tens of subconcepts as well as *genome component*. The query yields 1,031 relevant answers, but documents about specific stresses, such as *phosphate starvation* (97) were missed because *stress* is not defined in MeSH despite of its importance in Biology and it is then searched without any specialization. Five more documents could have been found if Entrez PubMed had lemmatized the documents. When MeSH term index is disabled, no document is retrieved as no paper full-text contains all the query words.

BioAlvis retrieved fewer documents than Google Scholar for two reasons: (1) its document collection is smaller (2) Google Scholar indexes scientific papers full text whereas BioAlvis only indexes abstracts and titles. It does not question the semantic annotation approach but the document collection preparation. BioAlvis missed also relevant documents because of the lack of some relevant sub-concepts of *stress* in the ontology like *acid shock*. This can be addressed by completing the ontology from a larger training corpus.

Regarding relevance of the documents, the accuracy of the answer sets varies a lot among the services. Google and Google Scholar results contain a vast amount of false positives. This is mainly due to the fact that the answer set contains many documents that mention only a subset of the query terms. The rank of these documents is very low but they are however counted in the answer set. The amount of false positives in Google Scholar is less important because the indexed corpus is smaller and more focused. Beyond the main problem of spurious co-occurrence of the query words mentioned above, the indexing of irrelevant subparts of the document caused many errors. For instance, citations of the document as occurring in Citeseer or SpringerLink sites are indexed with the document itself. BioAlvis retrieves false positives to a much lesser extent. Most of the irrelevant documents were papers about organisms other than *Enterobacteriacea*, many of them including a mention of a homology with the extensively studied enterobacterium *Escherichia coli.* This observation stresses the importance of filtering semantic annotations for IR purposes, so that semantic annotation focuses on the main topics of the paper.

5 Conclusion

While formal languages for ontology representation have made great advances, there are few formal or operational proposals designed to tie ontologies to linguistic knowledge [35]. Ontologies can no longer be considered as organized vocabularies or hierarchies of terms that can be simply mapped to the text for semantic markup. Intermediate linguistic knowledge levels are necessary to connect the textual information to the conceptual knowledge. Sophisticated and operational NLP platforms such as Alvis are available for developing such integrated applications. Still, the cost of maintaining and configuring them exponentially increases with the complexity of the linguistic knowledge. As highlighted above, the linguistic knowledge is scattered into various heterogeneous resources in order to feed distinct successive linguistic analyses.

In this paper, we pointed out the challenge of integrating ontology knowledge and linguistic knowledge into a consistent model. In order to alleviate the lack of specialized knowledge to feed NLP tools, knowledge acquisition and ML methods are applied to training corpora. This raises the problem of integrating the processes of knowledge resource acquisition and the exploitation of these resources. We proposed an operational approach based on the clear specification of the learning task and the normalization of the example representation. Following these principles, we developed large resources in Biology for each linguistic step and demonstrated their efficiency through the semantic annotation of a representative Web corpus and its use in an IR application [36].

Acknowledgment. This research has been partially supported by the RNTL ExtraPloDocs and the FP6-STREP Alvis projects. Special thanks are due to Philippe Bessières. Without his expert participation in Biology, BioAlvis development would not have been possible. The terminology part of this work has greatly benefited from Annick Lacombe’s expertise.

References

1. Entrez PubMed. http://www.ncbi.nlm.nih.gov/sites/entrez
2. Buitelaar P., Declerck T., Sacaleanu B., Vintar S., Raileanu D., Crispi C. A Multi-Layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations. In Proceedings of EACL NLPXML’03 Workshop Budapest, Hungary, 2003.
3. Fensel, D., Hendler, J. A., Lieberman, H. and Wahlster, W. (Eds.) Spinning the Semantic Web: bringing the World Wide Web to its full potential. Cambridge, MA: MIT Press, 2003
4. Bontcheva K., Tablan V., Maynard D., Cunningham H. Evolving GATE to Meet New Challenges in Language Engineering. Natural Language Engineering. 10:349-373. 2004.
5. IBM Unstructured Information Management Architecture http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.index.html
6. Nédellec, C., Nazarenko, A. and Bossy R. Ontology and Information Extraction. In Ontology Handbook., S. Staab, R. Studer (eds.), Springer Verlag, to appear, 2008.
7. Collier N. and Takeuchi K. Comparison of character-level and part of speech features for name recognition in biomedical texts. J. of Biomedical Informatics 37, 423-435, 2004.
8. Yeh A., Morgan A., Colosimo M., Hirschman L. BioCreAtIvE Task 1A: gene mention finding evaluation. BMC Bioinformatics 2005, 6(Suppl 1)
9. Marquez L., Padro L., Rodriguez H. A machine learning approach to POS tagging in Machine Learning Journal, Vol 39, Iss 1, pp 59-91, 2000.
10. Faure D. and Nédellec C. A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In Adapting lexical and corpus resources to sublanguages and applications, workshop of the 1st LREC, p. 1-8, Velardi P. (Ed.), Grenada, Spain, 1998.
11. Buitelaar P., Cimiano P., Loos B. Proceedings of the Workshop on Ontology Learning and Population at the 16th ECAI,□Valencia, Spain, 2004.
12. W3C Semantic Web Health Care and Life Sciences Interest Group. http://www.w3.org/2001/sw/hcls/
13. Alvis project. http://cosco.hiit.fi/search/alvis.html

14. Hamon T., Nazarenko A., Poibeau T., Aubin S., Derivière J. A Robust Linguistic Platform for Efficient and Domain specific Web Content Analysis. Proceedings of RIAO, Pittsburgh, 2007.
15. Zebra sofware. http://www.indexdata.dk/zebra/
16. Kripke S. A. Naming and Necessity. In Semantics of Natural Language. D. Davidson, G. Harman (eds.), Reidel, Dordrecht, pp. 253-355, 762-769, 1972.
17. Witten I. H., Frank E. Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco, 2005.
18. Kim J.-D, Ohta T. Tsuruoka Y., Tateisi Y. and Collier N. Introduction to the Bio-Entity Recognition Task at JNLPBA, Collier et al. (eds), Proc. of NLPBA/Coling wshp, 2004.
19. Ando R. K. BioCreative II Gene Mention Tagging System at IBM Watson. Proceedings of the Second BioCreative Challenge Evaluation Workshop. 2007.
20. Nédellec, C., Bessières, P., Bossy, R., Kotoujansky, A. and Manine, A.-P., Annotation Guidelines for Machine Learning-Based Named Entity Recognition in Microbiology. In Proceedings of the ECML/PKDD workshop Data and Text Mining in Integrative Biology. M. Hilario and C. Nedellec (eds), 40-54, 2006.
21. Alexa T., Mccray, Allen C., Browne, Bodenreider O. The lexical properties of the Gene Ontology. In Proceedings of AMIA Symposium, San Antonio, 2002.
22. MeSH thesaurus. http://www.nlm.nih.gov/mesh/
23. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29, 2000.
24. Aubin, S. and Hamon T.: Improving Term Extraction with Terminological Resources. In Proceedings of FinTAL'2006. pp. 380-387.
25. Jacquemin, C. A Symbolic and Surgical Acquisition of terms Through Variation. In Connectionist, Statistical and Symbolic Approaches to Learning for NLP, Wermter, S., Riloff, E. & Scheler, G. (eds), pp. 425-438, Springer-Verlag, 1996.
26. Pyysalo S., Salakoski T., Aubin S., and Nazarenko A. Lexical adaptation of Link Grammar to the biomedical sublanguage: a comparative evaluation of three approaches. BMC Bioinformatics, 7(Suppl 3), 2006.
27. Aubin S., Nazarenko A. and Nédellec C. Adapting a General Parser to a Sublanguage, Proceedings of RANLP'05, pp 89-93. Borovets, Bulgarie, 2005.
28. Bisson G., Nédellec C. et Canamero D. "Designing clustering methods for ontology building - The Mo'K workbench" in Proc. of the workshop on Ontology Learning, (ECAI-2000), Staab S. et al (Eds)., p. 13-19, Berlin, 2000.
29. Manine A.-P., Alphonse E. and Bessières Ph. Genic Interaction Extraction by Reasoning on an Ontology. In SMBM’2008, pp. 93-100, 2008.
30. Nédellec C., Ould Abdel Vetah M., and Bessières P. Sentence Filtering for Information Extraction in Genomics, a Classification Problem. In PKDD'2001, p. 326-338, 2001.
31. LLL dataset. http://genome.jouy.inra.fr/texte/LLLchallenge/
32. Nédellec C. Genic Interaction Extraction Challenge. In Proc. of the Learning Language in Logic (LLL05) workshop joint to ICML'05. Cussens J. and Nedellec C. (eds). Bonn, 2005.
33. Krallinger, M. The interaction-Pair and Interaction Method Sub-Task evaluation. In proceedings of the BioCreAtIvE II Workshop, 2007.
34. Buntine, W., Zhou, L., Toan, V. L., Hamon, T., Ardö, A., Nazarenko, A., Nedellec, C., Pedersen, G. and Podnar, Y. Report on Tests, D8.3 IST-FP6 Alvis project, 2007.
35. Buitelaar P., Declerck T., Frank A., Racioppa S., Kiesel M., Sintek M., Engel R., Romanelli M., Sonntag D., Loos B., Micelli V., Porzel R., Cimiano P. LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. In Proc. of OntoLex06, a Workshop at LREC, Genoa, Italy, 2006.
36. http://genome.jouy.inra.fr/alvis/front