Toward a Standard Lexical Resource in the Bio Domain

Valeria Quochi, Riccardo Del Gratta, Eva Sassolini, Monica Monachini, Nicoletta Calzolari

Istituto di Linguistica Computazionale Area della Ricerca CNR, Pisa, Italy {firstname.name}@ilc.cnr.it

Abstract

The present paper describes the model and database structure of a large-scale lexical resource for the biology domain designed both for human and for machine use. Our lexicon aims at semantic interoperability and extendability. This is achieved through the adoption of the up-coming ISO standard for lexical representation and through a granular and distributed encoding of relevant information. The first part of this contribution focuses on two aspects of the model that are of particular interest to the Biology community: the treatment of term variants and the alignment with ontological information. The second part of the paper describes the physical implementation of the model: a relational database equipped with a set of automatic uploading procedures. The population relies on an XML input data structure that allows for the structuring of terms with their related properties and to automatically "pull-and-push" them into the database. Peculiarity of the BioLexicon is that it combines features of both terminologies and lexicons: it is able to represent terms and their variants, with some of the semantic relations that link them, and also morphological, syntactic and lexical semantic properties of terms as lexical items.

1. Introduction

Due to the increasing production of literature in the biomedical field, intensive research is being carried out around the globe to develop language technologies to access this large body of literature and to extract knowledge from it, in order to make it easily available to researchers, students and other domain users (e.g. protein and gene databases like Uniprot and EntrezGene). Most of the resources available, however, are created mainly for human use, which makes them often not particulary useful for text mining and information extraction applications. Recently, efforts have been directed to the creation of large-scale terminological resources that merge information contained in various smaller resources: large thesauri based on a normalized nomenclature (Kors et al., 2005), extensible lexical and terminological databases like Termino (Harkema et al., 2004) and the SPECIALIST Lexicon (NLM, 2007). Access to and interoperability of biological databases, however, is still hampered, due to lack of uniformity and harmonization of both formats and information encoded. The current challenge in bioinformatics is to construct a comprehensive and incremental resource which integrates bio-terms encoded in existing different databases and which encodes all relevant properties according to the most accredited standards for the representation of lexical, terminological and conceptual information (Hahn and Markó, 2001). Our paper describes both the conceptual and the physical model of a large-scale lexical and terminological lexicon for biology (the BioLexicon) that is currently under development within the European BOOTStrep project1. The resource we describe learns from the state-of-the-art resources, esp. from the SPECIALIST Lexicon and Termino, and builds on our experience in the standardization and construction of lexical resources. The goal is to propose a standard for the representation of lexicons in the Bio domain, which could be finally also interoperable with other domain lexicons.

2. Related Works

In this section we briefly review two advanced stateof-the-art lexicons in the bio-medical domain: the UMLS SPECIALIST Lexicon and Termino.

The UMLS SPECIALIST Lexicon has been developed by the NLM as a wide coverage lexical basis for UMLS NLP tools (NLM, 2007). It encodes words gathered from texts and for each word a set of morpho-syntactic and lexical semantic information is specified (for example part-of-speech, complement pattern, etc.). The format, however, is designed to be optimal for its use by specific applications, so it may not be easily reusable nor interoperable with other resources and/or tools.

Termino, in its turn, has a more flexible structure: each type of information (e.g. POS, external sources, and others) is encoded in separate tables, so that the information can be combined in different ways according to the specific user needs (Harkema et al., 2004). The model, however, seems not to conform explicitly to any established lexical standard.

The design of the BioLexicon structure, presented in section 3. below, aims at merging the two main features of those two resources: richness of linguistic information encoded and modularity and flexibility. The major novelty of our proposal is the adherence to the upcoming ISO standards for lexicon representation (see below) together with a detailed model suited to represent sophisticated syntactic and semantic information. Additionally, the BioLexicon encodes part of the conceptual information typically contained in ontologies or thesauri: i.e. semantic relations such as *is_a* and *is_part_of*.

¹BOOTStrep (Bootstrapping Of Ontologies and Terminologies STrategic Project) is a Specific Targeted Research Project of the European Unions 6th Framework Programme within IST call 4. Six partners from four European countries (Germany, U.K., Italy, France) and one Asian partner from Singapore are involved in the project. www.bootstrep.eu

3. The BioLexicon Resource

The BioLexicon is a large-scale resource that combines terminological data coming from the various available databases (mostly UniProt, Swiss-Prot, ChEBI, the BioThesaurus and the NCBI taxonomy) enriched with lexical information extracted from texts. Morphological, syntactic and lexical semantic properties of terms will be represented for each term and term variant. The BioLexicon is conceived as a semantically-enriched lexicon encoding relevant information for the biology domain and designed to be flexible enough to adapt to different application needs: e.g. text mining, information extraction, information retrieval, multilingual access. Moreover, since we aim at semantic interoperability in the biology community, the ISO Lexical Markup Framework (Francopulo et al., 2006a) was chosen as the reference meta-model for the structure of the BioLexicon and the ISO Data Categories as the main building blocks for the representation of the entries. The lexical resource model that we present here accounts for (English) biological terms and lexical items and represents their morphological, syntactic and lexical semantic properties. The BioLexicon, therefore, combines both lexical items typically contained in general purpose lexicons and specific terms of the biology domain and encodes those linguistic pieces of information that domain ontologies typically lack, but which are crucial to improve text mining results (e.g. part-of-speech and subcategorization frames). Finally, both the conceptual model and especially its physical implementation, are tailored to the automatic population and enhancement of the resource, independently of the various native data formats.

3.1. The Model of the Lexicon

The BioLexicon is modeled into an XML DTD according to the LMF core model plus objects taken from the three NLP extensions for the representation of morphological, syntactic and (lexical) semantics properties of terms (http://lirics.loria.fr/doc_pub/LMF_revision_14.pdf). The set of lexical objects and relations can be seen as the skeleton of the lexical entries. The content that allows for their actual representation is formed by the set of Data Categories, i.e. features used to decorate such objects. Data Categories are a fundamental component of lexical description, which ensure the standardization of the resource. They represent pieces of linguistic content that are not part of the model proper, but are used to describe the various instances of its objects. To do this, it will only be necessary to refer to their identifier in the the standard, or user defined, repositories. The set of Data Categories to be used in the BioLexicon is created both by drawing them from the standard sets of the ISO Data Category Registry (Francopulo et al., 2006b; Wright, 2004; ISO-12620, 2006) and by creating specific ones for the Bio domain. In doing this, we also aim at establishing a standard set of Data Categories for this domain. Examples of ISO data categories are partOfSpeech (with its possible values) and written-Form. sourceDC, isRegulatorOf, and confidenceScore, instead, are examples of DCs specifically created for the biology domain on the basis of expert requirements. Fig. 1 shows the overall model of the BioLexicon. For reasons

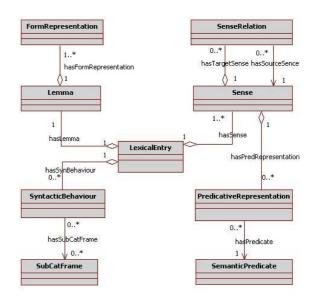


Figure 1: The BioLexicon model

of space, in this paper we focus only on its main lexical objects: *LexicalEntry, Lemma, Syntactic Behaviour* and *Sense*. The last three can be viewed as the central classes of three distinct modules for morphology, syntax and semantics.

Lexical Entry represents the abstract lexical unit, i.e. the lexeme; it contains information about the external resources/corpora from which terms are extracted and has the main function of linking all the three main levels of description: morphology, syntax and semantics. Following the philosophy of modularity and extensibility of the LMF, in the BioLexicon we represent external references as (typed) data categories that are added as attributes of the Lexical Entry object². One Lexical Entry, for example, will point to a specific Data Category typed as SourceDC containing the name of the original resource and the id of the specific term. This allows to add new sources and new IDs easily even to already existing Lexical Entries. **Lemma** represents the base form of the lexemes (i.e. the conventional forms of terms in the domain) plus additional grammatical features such as grammatical gender. The Lemma is in a one-to-one relation with the Lexical Entry, which forces homonyms to be represented as distinct lexical entries. Syntactic Behaviour is the object dedicated to the representation of syntactic units, i.e. how lexical items and terms are used in context. One Syntactic Behavior describes one specific set of syntactic properties of a lexical item: it specifies syntactic features and subcategorization frames of lexical items. The architecture of the syntax module is designed in view of accommodating the subcategorization behaviors of terminological verbs, and possibly nominalizations, automatically extracted from texts with appropriate NLP algorithms. Finally, Sense is the class used for the representation of the lexical meanings

²For consistency, as well as for practical, reasons we represent most of the content of the lexical objects as typed Data Categories.

of a word/term, and it is inspired by the SIMPLE Semantic Unit (Ruimy et al., 2003). Sense represents lexical items as semantic units. Each Sense instance represents and describes one meaning, contains information on the specific (sub)domain to which the term sense applies, and indicates the semantic type that it instantiates. This way we realize a first mapping to the BioOntology, in that the Semantic-Type DCs will map to ontological nodes represented in the BioOntology. By means of the other classes of the semantic module, the BioLexicon will also represent the semantic argument structure of lexical items, where appropriate, and semantic relations between senses. In particular, the SemanticPredicate and, indirectly, the SenseRelation classes (see fig. 1) are responsible for the mapping and alignment with the *BioOntology*, a domain ontology developed in parallel within BOOTStrep (by other partners and domain experts).

3.2. Snapshot I: the Treatment of Term Variants

In a terminological lexicon for biology a key requirement is the representation of the different types of term variants. Variants, in fact, are extremely frequent and common in the literature. (Nenadic et al., 2004), for example, distinguish different types of variants: orthographic variants, acronyms and abbreviations, morphological variants, structural variants, semantic variants. In the BioLexicon we chose to distinguish only between two types of variants, since this is what can be classified and extracted automatically by BOOTStrep partners: variants of form (i.e. orthographic variants, acronyms and abbreviations, and morphological variants) and variants of meaning. These two "macro"-types of variants are accounted for by means of two different representation devices: variants of form are represented in the FormRepresentation object, whereas semantic variants are represented as distinct but synonymous Lexical Entries. FormRepresentation has the function of representing multiple orthographies of terms: orthographic variants of the same lemma. The basic DC specifying the FormRepresentation is the writtenform, i.e. the string identifying the form in question. Each variant is adorned with properties represented by specific DCs: the type of variants ("orthographic", "acronym", ..., and "baseform"), and a confidenceScore assigned to each variant. The latter is a fundamental Data Category for the representation of term variants extracted from texts using machine learning algorithms in that it can be used to keep track of the "goodness" of each variant. Since the form of the lemma is also represented in as a FormRepresentation, this object can act as the bottom-up access to the whole lexicon. Fig. 2 is an XML excerpt that shows how variants (and SourceIds) are represented in the BioLexicon.

Semantic Variants of terms are treated differently. They are represented in the Semantic module as synonymous terms. The **SenseRelation** class fulfils this purpose: semantic variants are represented as distinct lexical entries and consequently as distinct senses, but are linked together by a synonymy relation in the semantic layer. An isSynonymOf relation between Senses is used to keep track of all semantic variants of a term entry.

```
<LexicalEntry ID="LE_interleukin-2">
<POSDC POSAtt="Pos" POSVal="Noun"/>
<SOURCEDC SourceAtt="UNIPROT
SourceVal = "11111"/>
<Lemma ID="LM_interleukin-2"</pre>
basename="interleukin-2" >
<DC att="grammgender" val="Feminine"/>
<FormRepresentation ID="RF_interleukin-2"</pre>
writtenform="interleukin-2">
<VariantDC VariantAtt="VariantType"</pre>
VariantVal="FullForm"/>
</FormRepresentation>
<FormRepresentation ID="RF_IL2"</pre>
writtenform="IL2">
<VariantDC VariantAtt="VariantType"</pre>
VariantVal="Acronym"/>
</FormRepresentation>
< /Lemma>
</LexicalEntry>
```

Figure 2: XML sample bio lexical entry

3.3. Snapshot II: the Alignment with the *BioOntology*

There are two main mechanisms in the BioLexicon to align its entries with a domain expert ontology: one is a kind of natural link between event concepts and the SemanticPredicate object, which represents an event meaning together with its associated semantic "arguments". A Semantic Predicate can inherited from the BioOntology and "translated" into the representation language of the lexicon (i.e. Qualia relations, see below). It represents a meaning that may be shared by more instances of the Sense class that are not necessarily considered as synonyms, for example because they do not belong to the same lexical class. A typical example is a verb and the corresponding nominal predicate (like regulate and regulation). Second, SenseRelation is another key mechanism that further realizes the mapping between the lexicon and the ontology within the project by an automatic derivation of a set of Qualia relations (Pustejovsky, 1995) from event concepts in the Ontology. In their turn, event concepts are translated into Semantic Predicates to which the derived relations are constrained (cfr. Self-reference). Specific Semantic Relations in the BioLexicon, again, are represented as Data Categories partly drawn from the standard Data Category Registry, and partly defined specifically for the Bio domain.

4. The BioLexicon DataBase

The software implementation of the BioLexicon consists of two modules: a relational database MySQL and a java-based loading software for the automatic population of the database. External to the DB, but fundamental for its automatic population, is an XML Interchange Format (XIF hereafter) specifically tailored to the BioLexicon structure (see section 4.2.).

4.1. Database Architecture

The database is structured into three logically distinct but strongly interconnected layers (see Fig. 3). The TAR-GET FRAME layer contains the actual BioLexicon tables, i.e. tables that directly instantiate the lexical objects and

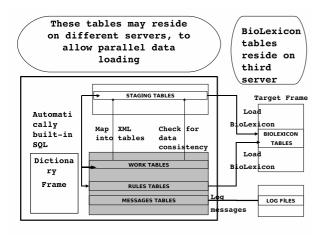


Figure 3: BioLexicon DB Architecture

relations designed in the conceptual model presented in the sections above and defined in a corresponding DTD. Each module of the BioLexicon structure (syntax, semantics, morphology) is independently accessible by queries and represents a self-consistent set of data. Each table has its internal identifier, but native identifiers, i.e. original source identifiers, are maintained as attributes of a given lexical entry (i.e. as DCs). The other two layers, DICTIO-NARY and STAGING, may be considered as operational layers: the DICTIONARY contains rules to populate target tables, the whereas STAGING is an intermediate set of hybrid tables for the storage of volatile data: staging table columns consist of attributes of the XIF and attributes of target tables. In addition, the staging frame is the level dedicated to data-cleaning and normalization. This neat separation between target tables (the BioLexicon proper) and operational tables allows for the optimization of the uploading of data into the BioLexicon DB and ensures extensibility both of the database and of the uploading procedures. Faced with the need to add a new table (i.e. a new lexical object) or a new attribute to an existing table, for example, it is sufficient to add only the definition of the new tables or attributes to the DB and to modify accordingly only the relevant portion of the Dictionary Frame layer to handle the novelties.

4.2. The XIF and Automatic Population

The interchange format (XIF) is designed with the main purpose of automatically populating the BioLexicon with data provided by domain experts and by lexical acquisition systems. Within the project, data are extracted and gathered through automatic procedures both from existing resources and by research papers in biology. The XML Interchange Format DTD is to be considered a simplified version of the BioLexicon DTD, which accommodates the needs of data providers and facilitates the automatic uploading of the DB. By means of the XIF, we therefore allow for a standardization of the data extracted from the different terminological resources and from texts. Differently from other similar lexical systems (like Termino, see Harkema et al. 2004), the XIF allows for the independency

of the uploading procedures from native data formats. This way, any system/group wishing to feed new data into the BioLexicon would only need to encode it in an XML file according to the XIF DTD. The XIF DTD partially mirrors the way biological data are stored into domain knowledge databases and also accommodates the way these data are extracted from those resources. The XIF is organized in clusters of terms, i.e. in sets of coherent types of information. A cluster contains one or more synonymous entries with information related to their lemmas, parts-of-speech, inflected forms, semantic relations and external references. Such an organization also permits the splitting of the input file by clusters, which in turn allows for a parallel uploading of the data into the DB. From the point of view of implementation, the XIF may be considered as the physical counterpart of the Dictionary Frame: that is to say, the loading software uses rules contained in the dictionary tables to correctly interpret the input file. Faced with the need to add a new table or to alter an existing one, it is sufficient to add new elements or new attributes to the XIF DTD and to add relevant instructions to the dictionary frame. The loading software interprets the XIF as usual and applies the new Dictionary instructions automatically inserting the new table or attribute. This property of the XIF together with the neat separation of the three layers of the DB mentioned above allows any agent (human or machine) to easily populate, update and create new objects and attributes. Recently, the first uploading of the BioLexicon DB has been completed: the data that are currently uploaded are terms gathered by our BOOTStrep partners from existing databases with information relevant to their external source references, variants, lexical category and, to some degree, semantic relations.

5. General Statistics and Content Validation

Since we are dealing with the building of a lexical resource within an ongoing project, no evaluation is available yet. However, some kind of content validation can be made, taking into account the input resources and the documentation so far produced. For validation of the resource we readapt templates from the ELRA Validation Manual for Lexica 2.0 (Fersøe, 2004)³. Here, we report some general interesting information about the lexical resource together with figures describing its present content.

The BioLexicon, a monolingual English lexical resource, represents Bio-terms as well as general lexical items relevant to the bio-medical domain. Both nouns and verbs are represented: nouns covers a wide portion of the existing biological terminologies and come from the most used databases in the sector. Mainly, they represents terms denoting enzymes, chemical, species, genes and/or proteins, especially those relevant for the gene regulation topic. Verbs are also represented, but are very limited in number: for the moment only verbs relevant to the E. Coli

³Such validations are especially important given the automatic uploading of data provided according to the XIF by different groups. This way we can monitor both the state of advancement of the resource, specifically its size, and the formal correctness of the data uploaded

species have been included. For each entry the part-ofspeech is encoded together with the written form of both its lemma and its variants. Also some semantic relations are instantiated: synonymy, part-of, and a few other biological relations. Finally, The BioLexicon resource is implemented as a MySQL relational database that runs both under Linux and Windows systems. The database is shaped according to the model XML DTD, and therefore easily allows for XML outputs. So far, the DB has automatically uploaded all the input files provided by the bio-experts within the BOOTStrep project (EBI-Cambridge and MIB-Manchester), which gather and systematize biological terminology from the major online databases. Altogether the DB contains 25 million records and occupies ca. 1.7G of memory space. It consists of 1,309,496 Lexical Entries, Lemmas and Senses; 2,047,903 orthographic variants and 1,697,804 semantic relations. Detailed figures about the different types of entries and information covered by the DB are reported in tables 1 and 2 below.

POS	Sem.Type	Lexical Entries	Variants
N	Enzymes	4,016	9,379
N	Genes/Proteins	841,164	1,547856
N	Species	367,565	439,336
N	Chemicals	13437	51332
POS	Sem.Type	Lexical Entries	Infl. Forms
V		489	2435

Table 1: Lexical Entries

Relation Type	#
is_a	464,078
is_synonym_of	847,873
is_part_of	189
is_conjugate_base_of	637

Table 2: Semantic Relations between Senses (Entries)

6. Concluding Remarks

With biological literature continuously developing, large-scale terminological lexicons that can support text mining and information extraction applications could make the life of biologists much easier. The existence of a reusable bio-lexicon with sophisticated linguistic information and a corresponding bio-ontology will enable the bioinformatics community to develop information extraction tools of higher quality than before. The lexicon model described in this paper is designed so as to integrate both typical information provided by domain ontologies and typical linguistic information generally available in open-domain computational lexicons: terms and variants are encoded with their semantic information as well as with typical linguistic information such as parts-of-speech and subcategorization frames. The DB, as a reflection of the model, is a modular and flexible resource which can automatically upload new data and provide (XML) outputs by means of web services. The main contributions of the resource to the biomedical field are the compliance to the most recent ISO Standard and the implementation of automatic procedures for the uploading of data provided by different sources and groups. A brief description of the size and coverage of the resource has been given in section 5., based on the first uploaded data coming from our project partners. An evaluation of the resource, instead, is not applicable at the moment and will only be possible at a later stage.

In the next future the BioLexicon will be enhanced by uploading terms and variants extracted from texts by our partners, together with sophisticated linguistic information: i.e. confidence scores for variants, subcategorization frames, syntactic and semantic arguments. Finally, the DB will be integrated into a common framework of shared and distributed resources like the IBM UIMA middleware (Ferrucci and Lally, 2004). To this end, both a web interface for human access and a series of web services for machine use will be designed possibly making use of the XIF as the input/output format and thus profit from its modularity.

7. References

Ferrucci, D. and A. Lally, 2004. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engeneering*, 10(3-4):327–348.

Fersøe, H., 2004. Validation manual for lexica. Release 2.0, ELRA.

Francopulo, G., T. Declerk, M. Monachini, and L. Romary, 2006b. The relevance of standards for research infrastructure. In *Proceedings of the LREC 2006*.

Francopulo, G., L. Romary, M. Monachini, N. Calzolari, et al., 2006a. Lexical markup framework (lmf). In *Proceedings of the LREC 2006*.

Hahn, U. and K. Markó, 2001. Joint knowledge capture for grammars and ontologies. In *Proceedings of the 1st international conference on knoledge capture*.

Harkema, H., R. Gaizauskas, M. Hepple R. Angus, I. Roberts, N. Davis, Y. Guo, et al., 2004. A large scale terminology resource for biomedical text processing. In *BioLINK*, ACL 2004. ACL.

ISO-12620, 2006. Terminology and other content language resources- data categories- specifications of data categories and management of a data category registry for language resources. Iso/tc37/sc3/wg4, ISO.

Kors, J.A., M.J. Schuemie, B.J.A. Schijvenaars, M. Weeber, B. Mons, et al., 2005. Combination of genetic databases for improving itentification of genes and proteins in text. In *BioLINK* 2005.

Nenadic, G., S. Ananiadou, and J. McNaught, 2004. Enhancing automatic term recognition through term variation. In *Proceedings of the 20th Coling*, 2004.

NLM, 2007. Specialist lexicon and lexical tools. Umls release documentation, NLM. Http://www.nlm.nih.gov/pubs/factsheets/umlslex.html.

Pustejovsky, J., 1995. Generative Lexicon. MIT.

Ruimy, N., M. Monachini, E. Gola, N. Calzolari, M.C. Del Fiorentino, M. Ulivieri, and S. Rossi, 2003. A computational semantic lexicon of italian: Simple. *Linguistica Computazionale*, XVIII-XIX:821–864.

Wright, S.E., 2004. A global data catergory registry for interoperable language resources. In *LREC* 2004.