Semantic Annotation in the Alvis Project

Adeline Nazarenko, Claire Nédellec, Erick Alphonse, Sophie Aubin, Thierry Hamon, Alain-Pierre Manine

Semantic Annotation in the Alvis Project*

Adeline Nazarenko[2], Claire Nédellec[1], Erick Alphonse[1], Sophie Aubin[2], Thierry Hamon[2], Alain-Pierre Manine[1]

[1] Laboratoire Mathématique, Informatique et Génome (MIG), INRA, Domaine de Vilvert, 78352 F-Jouy-en-Josas cedex.

[2] Laboratoire d'Informatique de Paris-Nord (LIPN), Université Paris-Nord & CNRS, av. J.B. Clément, F-93430 Villetaneuse.

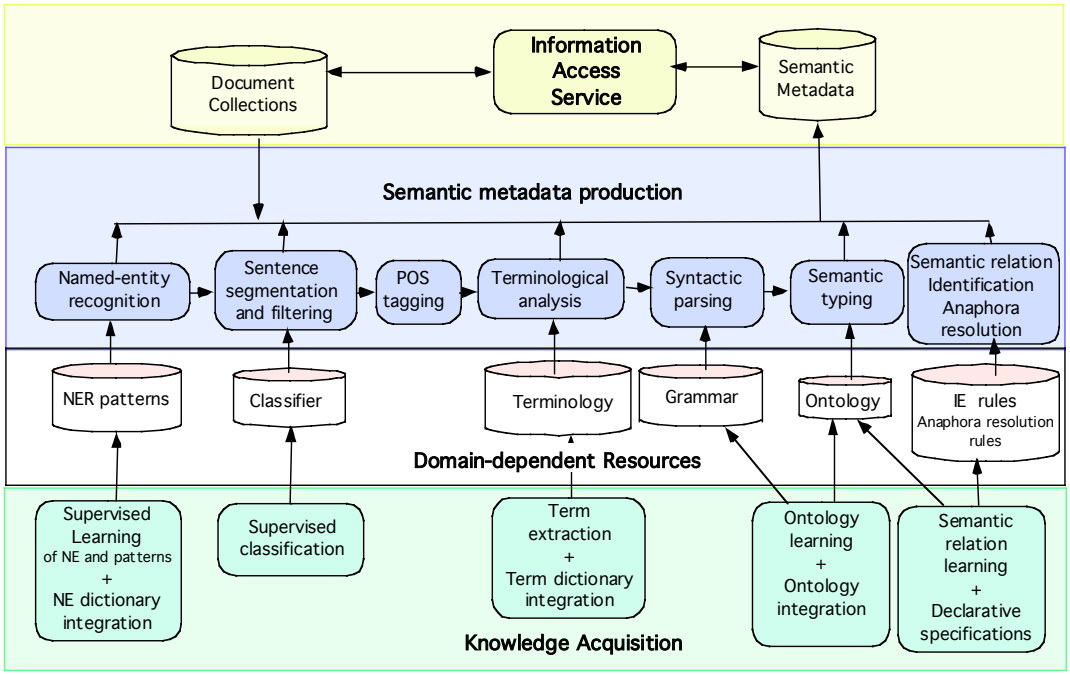
1 Introduction

In semantic search engines, the user query language includes constraints on the semantic types and relations instead of boolean constraints on simple words. For instance, the query

> Author=*person:Crick* and Author=*person:Watson* and Paper_title=*title:A structure of DNA* and Publication_date=*date:1953*

means that the user looks for documents including descriptions of a paper by Watson and Crick, published in 1953 and entitled "A structure of DNA ". Google-like queries based on simple keywords cannot express such roles and retrieve irrelevant papers where the cooccurrence of the four terms (Crick, Watson, DNA structure, 1953) is spurious. Semantic search also includes *term normalization* so that the term variations in the queries and in the documents are replaced by canonical terms. For instance "structure of DNA" is equivalent to "DNA structure" which can be considered as the canonical representant.

The development of such facilities requires an appropriate semantic annotation and indexing of the document collection. More generally, the poor quality of the semantic annotations of documents affects the development of new services of intelligent access to documents, including IE, Q/A or summarization. The semantic annotations should be more precise, more complete and non ambiguous in order to answer to the growing need of specific content access. Shallow processing of the documents such as word segmentation and stemming produce poor document representations in the form of sequences of words. It is not sufficient for identifying the relevant and canonical keywords and for associating them to the correct semantic types and relations. The main reasons are related to the homonymy (same word, different meanings) and synonymy (different words, same meaning) phenomena, to morpho-syntactic variations (*e.g.* noun1 of noun2 can be equivalent to noun2 noun1 in English), to anaphora (*e.g.* "it" stands for "Venus Express") and to the complexity of the sentences structures (*e.g.* the agent of the action may be far from the action). The meaning of words cannot be easily derived from the word itself and its neighborhood. More sophisticated linguistic processing is now recognized as needed for answering needs in specific domains such as retrieving relevant documents and extracting focused information [Ananiadou & Mc Naught, 2006].



* Published in the proceedings of the proceedings of the Int. Workshop on Intelligent Information Access (IIIA), July 6-8, 2006, Helsinki, Finland.

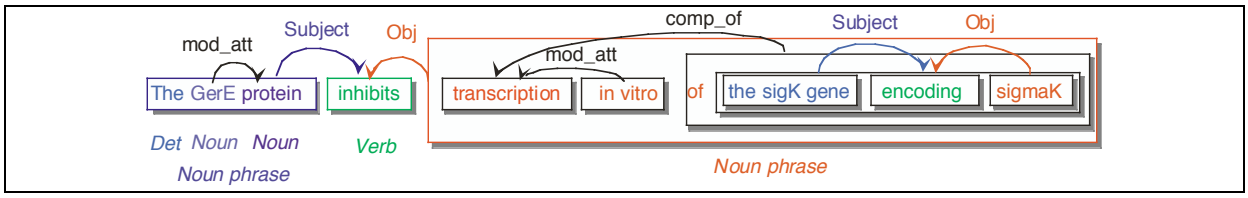
Figure 1: Semantic annotation of documents in Alvis

Alvis project[1] aims at developing a distributed semantic search engine for specific domains. The semantic indexing in Alvis relies on semantic annotations of words, terms and named-entities in the documents. Such a precise annotation requires a close integration of linguistic processing steps as presented in section 2 (in blue, Figure 1) and the exploitation of relevant linguistic resources, specific to the domain. Such resources are generally not available or insufficient with respect to the application needs in specific domains. Therefore Alvis project develops a machine learning-based platform for acquiring and adapting the resources such as named-entities, terminologies and ontologies (section 3) (in green, Figure 1).

We have chosen the microbiology domain as demonstration domain for three reasons: the availability of structured resources (*e.g.* document collections, nomenclatures), the clear expression of application needs such as the extraction of genic interactions for designing regulation networks and inducing functions [REFS] and the amount of previous work useful for evaluation and comparison purposes. Examples of this domain will thus illustrate the rest of the paper.

2 Linguistic processing

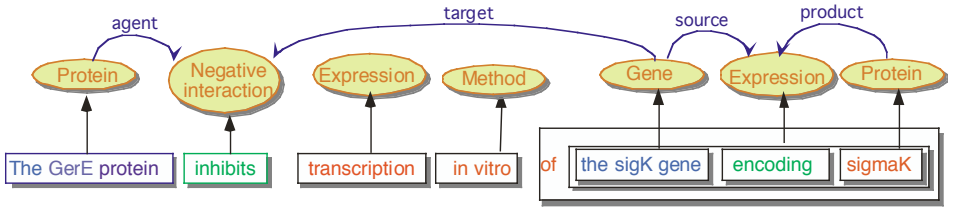
Assigning concepts to classes of documents can be done by applying statistics-based methods to shallow document representations such as bag of words and yields a sufficient precision for certain general purposes. However, semantic annotation of documents at the content level for domain-specific information retrieval, information extraction or question/answering cannot be done without a richer representation involving deep linguistic processing (Figure 2).



Syntactic dependency	subject(Ger_protein, inhibit), obj(transcription, inhibit), …
Terminology	term(GerE protein), term(in vitro) term(sigK gene)
Syntactic category	cat(the, det), cat(GerE, noun), cat(inhibit, verb), …
Named entity	entity(GerE), entity(sigK), entity(sigma K)
Word segmentation	word(the), word(Ger), word(protein), word(inhibit), …

Figure 2. Example of syntactic analysis

For readability reasons, the linguistic information is represented here in a logic-based language. Linguistic processing normalizes the document by replacing non-informative variations by canonical expressions such as stemming does. It also identifies linguistic units (words, terms, named-entities) and their relations (syntactic dependencies, anaphora). This linguistic analysis serves as a basis for the annotation of text units by concepts and semantic relations (Figure 3).



Semantic relation	agent(GerE protein,inhibit), target(sigK gene,inhibit), …
Semantic type	is_a(GerE protein, protein), is_a(inhibit, negative_interaction), …

Figure 3. Example of semantic annotation.

2.1 Linguistic platform

The linguistic analysis is performed by successive steps as illustrated in a simplified way by Figure 1. The actual execution order of the components is indeed more complex: for instance, a shallow NER step is applied for improving further sentence segmentation, while a full NER step is applied after morpho-syntactic tagging so that syntactic categories can be exploited for NER disambiguation. The relevant text can be filtered at any step on the basis of a "bag of word" representation of the text for focusing further processing.

[1] European Project STREP IST-1-002068-STP,http://www.alvis.info/alvis/

The objects to be identified are expressed by text units, namely the named-entities (NE) and the terms. A given object may be expressed by synonymous NE and terms and should be replaced by a same canonical form. Their processing raises different problems:

- Named-entity automatic and non-ambiguous recognition includes homonymy resolution (*e.g.* « cat » has more than seven different meanings in biology) and synonymy resolution (*e.g.* a given protein may have different names). These phenomena are so frequent that they greatly affect the document representation and weighting model for further document classification. In Alvis, the NE recognition is done through dictionaries and patterns involving typographic variations and context disambiguation as in [Tanabe & Wilbur, 2002]. In biology, NE recognition includes the attachment of proteins/genes to the relevant species.
- Terms recognition is done by using a terminology dictionary. Normalization exploits morpho-syntactic variation rules for associating the canonical form to variants, *e.g. human cancer* / *cancers in human*.

Once the object are all identified and roughly typed, they are then typed by finer-grain categories and relations. This requires syntactic dependency analysis that identifies the syntactic roles of the semantic units. It is useful both for disambiguating the semantic types and for deriving the semantic relations. We have adapted Link Parser as described and evaluated in [Aubin *et al.*, 2005].

- Semantic typing of semantic units (NE and terms) in Alvis is done by choosing the relevant concepts in the conceptual hierarchies of the ontology. Ambiguities are solved by verifying the semantic and syntactic constraints attached to the ontological concepts as described in [Faure & Nedellec, 1999]. The mapping between the ontological level and the text level is done through the lexical level. This level describes the linguistic expressions (NE and terms, their variations and canonical forms) to be mapped to the text and their links with the ontological knowledge.
- Domain specific relations in Alvis are identified between the objects of the domain (entities and terms) in the documents by Information Extraction rules (section 3).
- Anaphora and coreference resolution makes explicit relevant objects of the domain that are represented by pronouns, synonyms or properties by relating them to the full references (*e.g.* GerE = this protein = it).

The platform is designed in a modular way so that it can be reconfigured and the components can easily be replaced as in Gate philosophy [Cuningham *et al.*, 2002]. Each components of the NLP processing line takes as input and deliver as output, documents in XML format according to a layered DTD [Nazarenko *et al.*, 2006].

3 Learning

Accurate linguistic processing relies on specific resources. They can automatically be acquired from training corpus by machine learning and statistics-based methods that complete the existing lexicon [Nedellec & Nazarenko, 2005].

The main knowledge acquisition developments in Alvis project focuses on rule learning for named-entity recognition, term extraction, preference rule weighting for anaphora resolution, conceptual hierarchy learning for semantic typing and IE rule learning for semantic relation tagging (Figure 1). In the following, we will present in more details the acquisition of terms and of semantic relations. Our contribution to anaphora resolution can be found in [Weissenbacher, 2005]. Conceptual hierarchies are acquired by the Asium system as described in [Faure & Nedellec, 1999]. NE learning is ongoing work.

3.1 Term identification

YaTeA selects term candidates among the noun phrases that it has identified in thetraining corpus. It provides term syntactic analysis in a head-modifierformat. Its is able to efficiently process large corpora (EN dire PLUS sur les performances ou SUPPRIMER). YaTeA training corpus is preprocessed, segmented into words and sentences, lemmatized and tagged with part-of-speech (POS) information. Then, candidate term identification bu YaTeA mainly relies on recognition of predefined parsing patterns and endogenous disambiguation. Ambiguous attachments in noun phrases are solved by searching for non ambiguous sub noun phrases in the training corpus. For example, in *transcription sigma factor*, the attachment of *transcription* to *factor* and not to *sigma* can be done if the term *transcription factor* occurs in the corpus. Exogenous disambiguation of candidate terms is done by the use of external resources such as lists of testified terms. A noun phrase is considered as a termcandidate if at least one parsing pattern applies. This is performed in three main steps, corpus chunking (*i.e.* construction of alist of Maximal Noun Phrases (MNPs)), parsing (*i.e.* attempts to find at least one syntactic parsing for eachMNP) and, finally extraction of candidate terms. A preliminary experiment on a biomedical corpus has shown that existing terminologies has a positive effect on the quality of the identification of maximal noun phrases, on their parsing andon the extraction of lists of term candidates. (EN dire PLUS ou SUPPRIMER)

3.2 Learning rules for semantic relation tagging

Semantic relation tagging is done by applying rules that are learned from training corpus by the LP-Propal method. LP-Propal is based on the supervised ILP algorithm Propal [Alphonse & Rouveirol, 2000]. It takes a training corpus as input where the target relation is tagged by an expert with the XML editor Cadixe[2]. The corpus has to be fully processed by the linguistic line as represented in Figure 1. Then, given a declarative list of relevant syntactic dependencies and terms and conditions on their occurrence context, LP-Propal selects the relevant features of the training examples. For instance, in biology, the term *expression* can be removed when it occurs in "A activates the expression of B", because "A activates B" is fully equivalent. Removing such ellipsis reduces data sparseness and improves the training corpus homogeneity. LP-Propal learns multi-class rules such as,

Genic_interaction (X,Z):-

is-a(X,protein), subject(X,Y), cat(Y,verb), is-a(Y,interaction), obj(Z,Y), is-a(Z,gene-expression).

that means, If the subject X of an interaction verb Y is a protein name, and the object Z is a gene expression, then, X is the agent and Z is the target of the interaction. LP-Propal has learned this rule from the training datasets of the LLL challenge on genic interaction extraction[3]. It yields 42.8 % recall and 75 % precision on the action/without coreference subset, which is very promising with respect to the results of the other participants [Nedellec, 2005] and to comparable MUC event extraction challenges. The role of the syntactic parsing has been experimentally measured by applying LP-Propal to the same dataset with the neighborhood relation and without the syntactic dependencies. It yields a poor precision (22.8) and recall (34.7). Moreover, the abstraction step affects the recall (- 8 %) but dramatically augments the precision (+ 46 %).

4 Conclusion

Semantic search in Alvis project relies on a detailed semantic annotation of the documents. It is performed by a full linguistic analysis, focused on relevant extracts. The needed resources are automatically acquired by various ML-based methods applied on training corpora. Preliminary experiments on knowledge acquisition tasks yield encouraging results. The effect of using such resources for fine-grained semantic annotation has still to be measured by evaluating the performances of the search itself. We will in particular apply different linguistic processing and various resources for precisely measuring their impact.

References

Ananiadou, S. & McNaught, J. (2006) *Text Mining for Biology and Biomedicine*, pp. 1-12, Artech House Books.

Alphonse E. and Rouveirol C. (2000). Lazy propositionalisation for relational learning. In Horn W. (ed.). *14th European Conference on Artificial Intelligence (ECAI'2000),* Berlin, Germany, pp. 256-260, IOS Press.

Alphonse E., Aubin S., Bessières P., Bisson G., Hamon T., Lagarrigue S., Manine A.-P., Nazarenko A., Nédellec C., Ould Abdel Vetah M., Poibeau T., Weissenbacher D. (2004). Event-based information extraction for the biomedical domain: the Caderige project. In *Proceedings of International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, Ruch P., Collier N. and Nazarenko A. (eds.).

Aubin, S., Nazarenko, A. and Nedellec, C. (2005). Adapting a general parser to a sublanguage. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, 89-93, Borovets, Bulgaria.

Cunningham H., Maynard D., Bontcheva K., Tablan V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia.

Faure D. and Nédellec C. (1999). "Knowledge Acquisition of Predicate-Argument Structures from technical Texts using Machine Learning" in Proceedings of *Current Developments in Knowledge Acquisition: EKAW-99*, p. 329-334, Fensel D. et Studer R. (Ed.), Springer Verlag, Karlsruhe, Germany.

Nazarenko A., Alphonse E., Derivière J., Hamon T., Vauvert G., Weissenbacher D. (2006). "The ALVIS Format for Linguistically Annotated Documents" in *Proceedings of LREC'06*.

Nédellec C. (2005). "Learning Language in Logic - Genic Interaction Extraction Challenge" in Proceedings of the Learning Language in Logic (LLL05) workshop joint to ICML'05. Cussens J. and Nedellec C. (eds). Bonn.

Nédellec C. et 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.

Tanabe L, Wilbur WJ. (2002). *Tagging gene and protein names in biomedical text*. Bioinformatics 2002;18(8):1124--32.

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

[3] http://genome.jouy.inra.fr/texte/LLLchallenge/

Weissenbacher, Davy. (2005). "A Bayesian Network for the resolution of non-anaphoric pronoun *it*". In *Proceedings of the NIPS 2005 Workshop on Bayesian Methods for NLP Whistler, Canada*.