Book Title
Book Editors
IOS Press, 2003

Ontology and Information Extraction: synergy and cooperation

Claire Nédellec [a,1], and Adeline Nazarenko [b,2]

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

[b] *Laboratoire d'Informatique de Paris-Nord (LIPN), Université Paris-Nord & CNRS, 99 av. J.B. Clément, F-93430 Villetaneuse*

Abstract. Information extraction (IE) and ontologies are involved in two main and related tasks, which are combined in a cyclic process. IE needs ontologies as part of the understanding process for extracting the relevant information and IE extracts new knowledge from the text, to be integrated in an ontology. This paper shows that even in the simplest cases, IE is an ontology-driven process and that IE can be used to populate ontologies and structure ontological knowledge. This paper is illustrated in biology, a domain in which there are critical needs for content-based exploration of the scientific literature. It takes the example of the ExtraPloDocs project, which aims at extracting gene-protein interaction information from the bibliography in genomics.

Keywords. Information extraction, Ontology design, Machine Learning, Natural Language Processing, Extraction rules, Named entity recognition, Relation extraction,

1. Introduction

An ontology is a description of conceptual knowledge organized in a computer-based representation while information extraction (IE) is a method for analyzing texts expressing facts in natural language and extracting relevant pieces of information from these texts.

IE and ontologies are involved in two main and related tasks, which are combined in a cyclic process. IE needs ontologies as part of the understanding process for extracting the relevant information and IE extracts new knowledge from the text, to be integrated in an ontology. In this paper, we will argue that even in the simplest cases, IE is an ontology-driven process and we will show in which respect IE can be used to populate ontologies and structure ontological knowledge.

Extracting information from texts calls for lexical knowledge, grammars describing the specific syntax of the texts to be analyzed, as well as semantic and ontological knowledge. In this paper, we will not oppose the lexical and linguistic knowledge and the on-

[1]Correspondence to: Claire Nédellec, MIG-INRA, Domaine de Vilvert, F-78352 Jouy-en-Josas cedex; E-mail: claire.nedellec@jouy.inra.fr

[2]Correspondence to: Adeline Nazarenko, LIPN, Université Paris-Nord, 99, av. J.B. Clément, F-93430 Villetaneuse; E-mail: adeline.nazarenko@lipn.univ-paris13.fr

Document: Professor John Skvoretz, U. of South Carolina, Columbia, will present a seminar entitled "Embedded commitment", on Thursday, May 4th from 4-5:30 in PH 223D.

Form to fill (partial)	
place: ?	starting time: ?
title: ?	speaker: ?
Filled form (partial)	
place: PH 223D	starting time: 4 pm
title: Embedded commitment	speaker: Professor John Skvoretz

Figure 1. A seminar announcement event example.

tological one. We will rather consider ontologies as formal specifications of the domains of interest augmented with some part of linguistic knowledge. The ontologies that can be used for IE, and enriched by IE relate conceptual knowledge to its linguistic realizations (*e.g.* a concept must be associated with the terms that express it in the text, possibly in various languages).

This paper will be mainly illustrated in biology, a domain in which there are critical needs for content-based exploration of the scientific literature and that becomes a major application domain for IE. We will take here the example of the ExtraPloDocs project [18] in which the authors are involved. This project aims at extracting gene-protein interaction information from the bibliography in genomics.

2. Preliminaries

2.1. What is IE?

Developing intelligent tools and methods, which give access to document content and extract relevant information, is more than ever a key issue for knowledge and information management. IE is one of the main research fields that attempt to fulfill this need.

A typical IE task as defined the DARPA's MUC program (Message Understanding Conferences [43] is illustrated here by Fig. 1 from a CMU corpus of seminar announcements [22]). The IE process recognizes a name (*John Skvoretz*) and classifies it as a person name. It also recognizes a seminar event and creates a seminar event form (John Skvoretz is the seminar speaker whose presentation is entitled "Embedded commitment").

Even in such a simple example, IE should not be considered as a mere keyword filtering method. Filling a form with some extracted words and textual fragments involves a part of interpretation. Any fragment must be interpreted with respect to its "context" (*i.e.* domain knowledge or other pieces of information extracted from the same document). In the document of Fig. 1, "4-5:30" is understood as a time interval and background knowledge about seminars is necessary to interpret "4" as "4 pm" and as the seminar starting time.

Operationally, IE relies on document preprocessing and extraction rules (or extraction patterns) to identify and interpret the information to be extracted. The extraction rules specify the conditions that the preprocessed text must verify and how the relevant textual fragments can be interpreted to fill the forms. In the simplest case, the textual fragment and the coded information are the same and there are neither text preprocessing nor interpretation.

More precisely, in a typical IE system, three processing steps have been identified [30,17]:

- *Text preprocessing*, whose levels range from mere text segmentation into sentences and sentences into tokens to a full linguistic analysis;
- *Rule selection*: the extraction rules are associated with triggers (*e.g.* keywords), the text is scanned to identify the triggering items and the corresponding rules are selected;
- *Rule application* that checks the conditions of the selected rules and fills the forms according to the conclusions of the matching rules.

Experiments have been made with various kinds of rules, ranging from the simplest ones [53] (*e.g.* the subject of the passive form of the verb "murder" is interpreted as a victim) to sophisticated ones as in [64]. The more abstract (*e.g.* the more semantic and conceptual) the IE rule, the more powerful, concise and understandable it is. However, it requires the input text being syntactically parsed and semantically tagged in order to map to the rule abstract conditions. As shown in Fig. 2, the condition part of the extraction rules may check the presence of a given lexical item (*e.g.* the verb *named*), the syntactic category of words and their syntactic dependencies (*e.g.* object and subject relations). Different clues such as typographical characteristics, relative position of words, semantic tags[1] or even coreference relations can also be exploited. Most IE systems therefore involve linguistic text processing and knowledge: segmentation into words, morpho-syntactic tagging (the part-of-speech categories of words are identified), syntactic analysis (sentence constituents such as noun or verb phrases are identified and the structure of complex sentences is analyzed) and sometimes additional processing, such as lexical disambiguation, semantic tagging or anaphora resolution.

However, the role and the scope of the linguistic analysis differ from one IE system to another. Text analysis can be performed either as preprocessing or during extraction rule application. In the first IE systems [30], local and goal-driven analysis was preferred to full text preanalysis to increase efficiency, and the text preprocessing step was kept to minimum. Although costly, data-driven, full text analysis and normalization can improve the IE process in various manners. (1) It improves further NL processing steps, *e.g.* syntactic parsing improves attachment disambiguation [5] or coreference resolution. (2) Full text analysis and normalization also facilitates the discovery of lexical and linguistic regularities in specific documents. This idea, initially promoted by works on sublanguages [27,59] for tuning NL processing to a given type of texts, is now popularized by Machine Learning (ML) papers in the IE field for learning extraction rules. There are two main reasons for that. First, annotating training data is costly and the quantity of data to be annotated decreases with the normalization (the less variations in the data, the less data annotation is needed). Next, ML systems tend to learn non-understandable rules by picking details in training examples that do not seem to be related. Normalizing the text by representing it in a more abstract way increases the understandability of the learned rules. However, normalization also raises problems such as the biased choice of the right representation *before learning*, that is not dealt within the IE literature.

[1] *E.g.* If the verbs "named", "appointed" and "elected" of Fig. 2 were all known as 'nomination' verbs, the fourth condition of the rule could have been generalized to their semantic category 'nomination'.

Document: NORTH STONINGTON, Connecticut (Business Wire) - 12/2/94 - Joseph M. Marino and Richard P. Mitchell have been named senior vice president of Analysis & Technology Inc. (NASDAQ NMS: AATI), Gary P. Bennett, president and CEO, has announced.

Rule

Conditions:

noun-phrase (PNP, head(isa(person-name))),
noun-phrase (TNP, head(isa(title))),
noun-phrase (CNP, head(isa(company-name))),
verb-phrase (VP, type(passive),head(named or elected or appointed)),
preposition (PREP, head(of or at or by)),
subject (PNP, VP),
object (VP, TNP),
post_nominal_prep (TNG,PREP),
prep_object (PREP, CNP)

Conclusion:

management_appointment (M, person(PNP), title (TNP), company (CNP)).

Comment:

if there is a noun phrase (NP) whose head is a person name (PNP), an NP whose head is a title name (TNP), an NP whose head is a company name (CNP), a verb phrase whose head is a passive verb (named or elected or appointed), a preposition of, at or by,
if PNP and TNP are respectively subject and object of the verb,
and **if** CNP modifies TNP,
then it can be stated that the person "PNP" is named "TNP" of the company "CNP".

Labeled document

NORTH STONINGTON, Connecticut (Business Wire) - 12/2/94 - <Person>Joseph M. Marino and Richard P. Mitchell</Person> have been named <Title>senior vice president</Title> of <Company>Analysis & Technology Inc</Company>. (NASDAQ NMS: AATI), Gary P. Bennett, president and CEO, has announced.

Figure 2. Example from MUC-6, a newswire about management succession.

We will see in the following that these two approaches, in which text analysis is respectively used for interpretation (goal-driven) and normalization (data-driven), are very much tangled, as any normalization process involves a part of interpretation. One of the difficulties in designing IE systems is to set the limit between local and global analysis. Syntactic analysis or entity recognition can be performed on a local basis but are improved by knowledge inferred at a global level, because ambiguous cases of syntactic attachments or entity classification can be solved by comparison with non-ambiguous similar cases of the same document.

The MUC competition framework has gathered a large and stable IE community. It has also drawn the research towards easy to develop and efficient methods rather than strong and well-founded NLP theories. Semantic analysis is rather considered as a way to disambiguate the syntactic tagging and analysis than as a way to build a conceptual interpretation. Today, most of the IE systems that involve semantic analysis exploit the most simple part of the whole spectrum of domain and task knowledge, that is to say, named entities. However, the growing need for IE application to domains such as functional genomics that require more text understanding pushes towards more sophisticated seman-

INTERACTION:	Type:	negative, positive
	Agent:	any protein
	Target:	any gene

Figure 3. An example of IE form in the genomics domain, as a part of the biological model of gene regulation network, proteins interact positively or negatively with genes

tic knowledge resources and thus towards ontologies viewed as conceptual models, as it will be shown in this paper. The ExtraPlodocs project is based on this assumption.

2.2. *The role of ontologies in IE*

Even though ontologies usually do not appear as an autonomous component or resource in IE systems, we argue that IE relies on ontological knowledge.

An ontology identifies the entities that have a form of existence in a given domain and specifies their essential properties. It does not describe the spurious properties of these entities. On the contrary, the goal of IE is to extract factual knowledge to instantiate one or several predefined forms. The *structure* of the form (*e.g.* the example of genic interaction in Fig. 3) is a matter of ontology whereas the *values* of the filled template usually reflect factual knowledge (as shown in Fig. 1 above) that is not part of an ontology. In Sect. 3.4, we will show that IE is ontology-driven in that respect.

The status of the named entities is a pending question. Do they belong to an ontology or are they factual knowledge? In this paper, we will consider that entities, being *referential* entities, contribute to populate an ontology and, as such, are part of a domain ontology.

Whether one wants to use ontological knowledge to interpret natural language or to exploit written documents to create or update ontologies, in any case, an ontology has to be connected to linguistic phenomena. An ontology must be linguistically anchored. A large effort has been devoted in traditional IE systems based on local analysis to the definition of extraction rules that achieve this anchoring. In numerous IE applications the ontological knowledge is encoded as a keyword rule, which can be considered as a kind of compiled knowledge. In more powerful IE systems, the ontological knowledge is more explicitly stated in the rules that bridge the gap between the word level and text interpretation. As such, an ontology is not a purely conceptual model, it is a model associated to a domain-specific vocabulary and grammar. For instance, the rule of Fig. 2 above, states that a management appointment event can be expressed through three verbs (*named, elected or appointed*). In the IE framework, we consider that this vocabulary and grammar are part of an ontology, even when they are embodied in extraction rules.

The complexity of the linguistic anchoring of ontological knowledge is well known and should not be underestimated. A concept can be expressed by different terms and many words are ambiguous. Rhetorical phenomena, such as lexicalized metonymies or elisions, introduce conceptual shortcuts at the linguistic level that must be clarified to be interpreted into domain knowledge. A noun phrase (*e.g.* "the citizen") may refer to an instance (a previously mentioned specific citizen) or to the class (the set of all the citizens), thus leading to a very different interpretation. These phenomena, which illustrate the gap between the linguistic and the ontological levels, strongly affect IE performance. This explains why IE rules are so difficult to design.

IE is a targeted textual analysis process. The target information is described in the structure of the forms to fill. MUC has identified various types of forms describing elements or entities, events and scenarios. However, IE does not require a whole formal ontological system but only parts of it. We consider that the ontological knowledge involved in IE can be viewed as a set of interconnected and concept-centered descriptions, or "conceptual nodes"[2]. In conceptual nodes the concept properties and the relations between concepts are explicit. These conceptual nodes should be understood as chunks of a global knowledge model of the domain. The use of this type of knowledge in NLP systems is traditional [61] and is illustrated by MUC tasks.

Ontologies and IE are closely connected by a mutual contribution. An ontology is required for the IE interpreting process and IE provides methods for ontological knowledge acquisition. Even if using IE for extracting ontological knowledge is still rather marginal, it is gaining in importance. We distinguish both aspects respectively in the following sections, although we consider the whole process as a cyclic one. For instance, in the ExtraPloDocs approach, a first level of ontological knowledge (*e.g.* entities) helps to extract new pieces of knowledge from which more elaborated abstract ontological knowledge is designed, which in turn helps to extract new pieces of information in an iterative process.

3. Ontology for Information extraction

Since the template or form to be fulfilled by IE is a partial model of world knowledge, any IE system is ontology-driven. The ontological knowledge is primarily used for text interpretation. How poor the semantics underlying the form to fill may be, whether it is explicit [24,19] or not [22], IE is always based on a knowledge model. In this section, for exposition purposes, we distinguish different levels of ontological knowledge: the referential domain entities, the conceptual hierarchy, chunks of a domain model (i.e. conceptual nodes) and the domain model itself.

3.1. Sets of entities

Recognizing and classifying named entities in texts require knowledge on the domain entities. Specialized lexical or key-word lists are commonly used to identify the referential entities in documents. In the financial news of MUC-5, lists of company names have been used. In the context of cancer treatment, [56] makes use of the concepts of the Metathesaurus of UMLS to identify and classify biological entities (mostly proteins, genes and drugs). In different experiments, some lists of gene and protein names are exploited: [31] makes use of SWISS PROT protein list, whereas [47] combines pattern matching with a manually constructed dictionary. The machine learning based event extraction systems also usually make use of list of entities to identify the referential entities in documents [53,64,35,63,14]. The use of a lexicon and dictionaries is however controversial. Some authors like [42] argue that entity named recognition can be done without it.

[2]We define a conceptual node as a piece of ontological model to which linguistic information can be attached. It differs from the "conceptual nodes" of [64], which are extraction patterns describing a concept. We will see below that several extraction rules may be associated to a unique conceptual node.

At a first level, these lists of entities are used for semantic tagging. The entities (*e.g. Tony Bridge*) are actually described by their types (here PERSON) and by the list of the various textual forms that may refer to them[3] (*Mr. Bridge, Tony Bridge, T. Bridge*). However, exact character strings are often not reliable enough for a precise entity identification and semantic tagging[4]. In biology, for instance, some names like *2CAT* may have more than 10 different meanings. Then as highlighted by [63], providing the system with lists of entities does not help that much, "because too many of the relevant terms in the domain undergo shifts of meaning depending on context for simple lists of words to be useful". The connection between the ontological and the textual levels must then also rely on contextual rules, which are associated to named entities to help their identification and disambiguation.

As a by-effect, these resources are also used for naming normalization. For instance, the various forms of *Mr. Bridge* will be tagged as PERSON and associated with its canonical name form: <PERSON id=Tony Bridge>. Specialized genomics systems are particularly concerned with the variation problem, which introduces typographical alterations as well as very different synonyms when the naming nomenclature evolve. In Flybase[5], 40% of the gene names are associated with such synonyms. A large part of the research effort in IE to genomics has focused on the problem of identifying protein and gene names [49,23,16] and more recently, BioCreative challenge [68] and the NLPBABioNLP shared task [15]. In many cases, rules rely on shallow constraints rather than morpho-syntactic dependencies as presented in [45].

Beyond typographical normalization, ExtraPloDocs uses the semantic tagging of entities to normalize the sentences at a linguistic level. This tagging solves some syntactic ambiguities, for example if *cotA* is tagged as a *gene* in the sentence "the stimulation of cotA expression", knowing that a gene expresses proteins helps to understand that "cotA" is the agent of the expression rather than its patient. Semantic tagging is also traditionally used for anaphora resolution: [50] makes use of UMLS[6] types to identify and order the potential antecedents of an anaphoric pronoun (*it*) or noun phrase (*these enzymes*, *both genes*).

3.1.1. Hierarchies

Beyond the lists of entities that populate it, an ontology is formerly structured as a hierarchy of concepts. A hierarchy of semantic or word classes can be derived from this conceptual structure. Traditionally, IE focuses on the use of word classes rather than on the use of the hierarchical organization. For instance, in WordNet [39], the word classes (synsets) are used for the semantic tagging and disambiguation of words but the hyponymy relation that structures the synsets into a hierarchy of semantic or conceptual classes is seldom exploited for ontological generalization inference. The hierarchy should however help to design extraction rules with the proper level of abstraction.

Some ML-based experiments have been done to exploit hierarchies of WordNet and of specialized lexicons, such as UMLS [64,10,22]. The ML systems learn extraction rules by generalizing from annotated training examples. The difficult choice of the correct level in the hierarchy is left to the systems. Chai et al.'s system automatically learns for

[3]These various forms may be listed extensionally or intentionally by variation rules.
[4]In the above example, the string "Bridge" could also refer to a bridge named "Tony".
[5]http://flybase.bio.indiana.edu
[6]http://www.nlm.nih.gov/research/umls/

P-A structure of *activate*

Pred:	*activate*		Frame:	ACTIVATE	
	args:	subject (1)		slot:	agent (1)
		object (2)		slot:	target (2)

Figure 4. Example of a conceptual-node driven rule in functional genomics.

each relevant NP in the rule, the optimal level of semantic generalization on the WordNet hyperonym path by climbing WordNet hierarchies. For ambiguous words, which have several hyperonyms, the choice of the right hierarchy to climb is based on the user selection of the headword senses in a training corpus. Chai et al. argue that generalization along WordNet hierarchy brings a significant benefit to IE but that the incompleteness of WordNet in specific domains and the word sense ambiguity are important hindrances. The IE learning system, SRV, also uses semantic class information such as synsets and hyperonym links from WordNet lexicon to constrain the application of the IE rules, but [22] concludes that the improvement is not clear.

In specific domain such as genomics, the main problem is therefore the acquisition of domain dependent hierarchies. A lot of work has been devoted to their manual or automatic acquisition for a wide range of NL processing tasks in order to overcome the general ontologies limitations.

3.1.2. Conceptual nodes

The ontological knowledge is not always explicitly stated as it is in [24], which represents an ontology as a hierarchy of concepts, each concept being associated with an attribute-value structure, or in [19], which describes an ontology as database relational schema. However, ontological knowledge is reflected by the target form that IE must fill and which represents the *conceptual nodes* to be instantiated. Extraction rules ensure the mapping between a conceptual node and the potentially various linguistic phrasing expressing the relevant elements of information.

Most of the works aiming at extracting gene/protein interactions are based on such event conceptual nodes. In [69], predicate-argument structures (P-A structures), also referred as subcategorization frames, describe the number, type and syntactic construction of the predicate arguments. The P-A structures are used for extracting gene and protein interactions (see Fig. 4). The mapping between P-A structures and event frames (event conceptual nodes) is explicit and different P-A structures can be associated to a same event frame. For instance, the extraction of gene/protein interactions is viewed as the search for the subject and the object of an interaction verb, which are interpreted as the agent and the target of the interaction. These works rely on shallow, robust or full parsers, which do, or do not handle coordinates, anaphora, passive mood and nominalization [62,66,57,48,36,52]. Additional semantic constraints may be added as selectional restrictions[7] for disambiguation purposes. activate is an interaction verb

Considerable effort has been made towards designing automatic methods for learning extraction rules that map the syntactic categories, dependencies and semantic types into a conceptual node. An interesting example is the system RHB+ [60], which learns this mapping with the help of case-frames in Fillmore's sense [21]. RHB+ is able to com-

[7] A selectional restriction is a semantic type constraint that a given predicate enforces on its arguments.

bine multiple case-frames to map a unique conceptual node. The main difficulty arises from the complexity of the text representation once enriched by the multiple linguistic and conceptual levels. The more expressive the text representation, the larger is the search space for the IE rule and the more difficult the learning. The extreme alternative consists in either selecting the potentially relevant features before learning with the risk of excluding the solution from the search space, or leaving the system the entire choice, provided that there is enough representative and annotated data to find the relevant regularities. For instance, the former consists in normalizing by replacing names by category labels when the latter consists in tagging without removing the names. The learning complexity can even be increased when the conceptual or semantic classes are learned together with the conceptual node information as in [55,70].

3.1.3. Domain conceptual model

The link between the syntactic level and the event and scenario description is not always so straightforward. Beyond linguistic analysis [32,12], the text interpretation may require inference reasoning with domain knowledge. For instance, to be able to extract :

INTERACTION:	Type:	negative
	Agent:	sigma K
	Target:	spoIIID

from, "[...], such that production of sigma K leads to a decrease in the level of spoIIID.", more biological knowledge is necessary to interpret the protein level changes in term of interaction. P-A structures as those above will be useful at the lower level for interpreting the text and build a semantic structure but a causal model stating that correlation in protein quantity variations can be interpreted as an interaction is needed to connect and interpret the instantiated syntactic structures at a conceptual level.

3.1.4. ExtraPloDocs approach for extracting gene-protein interactions

The ExtraPloDocs project follows theses tracks and is heavily ontology-driven [2]. Extracting gene-protein interactions from the bibliography is a popular but challenging IE task since the bibliographic style is a complex one as shown in the following example:

> **GerE** *stimulates* **cotD** transcription and *inhibits* **cotA** transcription in vitro by sigma **K** RNA polymerase, as expected from in vivo studies, and, unexpectedly, profoundly *inhibits* in vitro transcription of the gene (**sigK**) that encode **sigma K**.

As the work mentioned above, we argued that extracting genic relations requires rich extraction rules [6] based at least on syntactic and semantic categories (*e.g. stimulates* is an interaction verb), on syntactic dependencies (*GerE* is the subject of *inhibits*) and the recognition of named entities (in bold in the example above). The originality here relies in the role of Machine Learning for acquiring the needed resources and the development of a whole Natural Language Pocessing line to normalize the original data, i.e. MedLine abstracts. The integration of these various processing steps raises new research problems that are not apparent otherwise.

As said above, in ExtraPloDocs, the recognition and normalization of named entities are based on genomic existing resources (GenBank, SwissProt) and state of the art methods (typographical variation and contextual pattern matching). A specialized hierarchy of semantic classes is used for disambiguating syntactic parsing and typing entities (*GerE* is-a GENE, *RNA polymerase* is-an ENZYME) and actions (*stimulates* is_an INTERACTION). The Asium software [20] is used to semi-automatically acquire these relevant semantic categories. It is based on an original ascendant hierarchical clustering method that builds a hierarchy of semantic classes from the syntactic dependencies parsed in the training corpus.

The extraction rules are thus applied on texts enriched with a lot of linguistic and ontological knowledge. They are themselves learned from a training corpus in which the interactions have been annotated. Learning IE rules is seen as usual a classification task, where the concept to learn is an n-ary relation between arguments, which correspond to the template fields. The learning algorithm is provided with a set of positive and negative examples of genic interactions built from the sentences annotated and linguistically normalized (which includes lemmatization, term recognition and syntactic dependency parsing). We use the relational learning algorithm, Propal [1]. On preliminary experiments, the performance of the learner evaluated by ten-fold cross-validation is 69 (6.5 %) of recall and 86 (3.2 %) of precision. This result is encouraging, showing that the normalization process provides a good representation for learning IE rules with both high recall and high precision[8]. For instance, the following learned IE rule:

genic-interaction (X, Z):- protein(X), gene(Z), interaction(X,V), subject(X,V),
obj(U,V), NprepN(of)(Z,U).

states that if X is the subject of an interaction verb V and a protein name and if the object of the verb is the expression of a gene Z, then X is the agent and Z the target of the interaction.

4. Information extraction for ontology design

Acquisition of ontological knowledge is a well-known bottleneck for many AI applications and a large amount of work has been devoted to knowledge acquisition from text. The underlying idea, inherited from Harris' work on the immunology sublanguage [28], is that, in specific domains, the linguistics reflects the domain conceptual organization. Even if the linguistic representation of the conceptual domain is biased, it remains one of the most promising approaches to knowledge acquisition. Following [38], a large amount of work has been devoted to term extraction [9,34] as a means to identify the concepts of a given domain and thus to bootstrap ontology design [26,44,4] (see also Ryu and Choi in this volume). Identifying how these terms relate to each other in texts helps to understand the properties and relationships of the underlying concepts.

Various methods are applied to corpora to achieve this acquisition process: endogenous distributional or cooccurrence analysis and rule-based extraction are complementary in this respect. We focus here on the latter approach, which pertains to IE. Reinberger and Spyns' chapter (in this volume) illustrates the former. We show that it can indeed contribute to the ontology acquisition and enrichment process. Rule-based extraction produces elementary results that are interpreted in terms of chunks of ontologi-

[8] The description of the IE task and the data including some linguistic information are available on the web page of the LLL'05 challenge [37].

cal knowledge: the referential entities and their interrelationships. Once extracted, these chunks have to be integrated into an ontology. We do not deal with that point here, as it goes beyond IE.

4.1. Entity name extraction

As explained in Sect. 2.2, we consider here that the referential entities (*e.g.* persons, dates or genes), which are usually represented as instances of concepts, are part of the ontology. In this perspective, there is a need for "populating" ontologies with the referential entities of the domain of interest by automatic ways; IE has also been widely used for the acquisition of this type knowledge. Extraction patterns are used to recognize and categorize previously unknown names of entities in documents, either specialized texts or web pages. The extraction methods differ regarding their pattern design technique, which is either automatic or manual.

Various methods have been tested to achieve automatic pattern learning. Hidden Markov Models (HMM) based on sequences of bigrammes (pairs of tokens) has become a popular method for learning named entity recognition patterns from annotated corpora [7] because simple bigrammes appear as sufficient for learning efficient rules. For instance, for the recognition of biological entity names, [16] relies on an HMM trained on 100 MedLine abstracts using only character features and lexical information. The results (F-score 73 %) are much better than those obtained by previous hand-coded patterns as reported by [23]. More recently, approaches based on the Maximum Entropy (ME) appear as very powerful and relevant [41,8,11]. As in HMM, the method computes the probability to output a given label, given the word to tag. In this model, dependencies between word labels are easier to represent and the role of useful text features[9] is more explicit and easier to take into account. Classical ML discriminant classification methods such as SVMs [65,33], k-KNN, Neural Networks have also been applied [71]. However, depending on the tasks and the type of entities, SVMs, ME and HMM yield more or less similar results.

While the pattern learning approach tends to use very basic information from the text, the hand-coded pattern approach relies more heavily on linguistics, external ontologies and context. The EDGAR system [56] identifies unknown gene names and cell lines by two ways: the concepts of UMLS and hand-coded contextual patterns, such as appositives, filtered through UMLS and an English dictionary and occurring after some signal words, (*e.g.* cell, clone and line for cells). A second phase identifies cell features, (*e.g.* organ type, cancer type and organism) by a similar mechanism. In [49] and [31], the recognition of gene and biological entity names relies on a combination of cues: grammatical tagging, contextual hand-coded patterns, specific lexicon (*e.g.* SWISS-PROT keyword list) and word morphological. The results obtained by [49] on a FlyBase corpus are of high quality, (94,4 % recall and 91,4 % precision). Populating ontologies with the help of entity name recognition from textual data can therefore be considered as operational for specific domains.

---

[9] Simple words, case, length, POS tags, semantic categories, numbers, specific symbols, prefix, suffix, context.

4.2. Relation extraction

In a structured ontology, the concepts are related to each other according to a variety of relations. Three main approaches acquire ontological relations from texts:

- The cooccurrence-based method identifies couples of cooccurring terms. When applied to large corpora, this method is robust but further interpretation is required to type the relation underlying the collocation.
- The knowledge-based method makes use of a bootstrapping dictionary, a thesaurus or an ontology and tunes it to adapt it to the specific domain at hand according to a representative "tuning" corpus.
- The IE pattern-based method.

The IE approach has the advantage over the first one that the type of extracted relation is known, since patterns are designed to characterize a given relation. It is complementary to the second one: preexisting knowledge can help to design an extraction rule in an acquisition iterative process. For instance, if the preexisting knowledge base states that 'X is-part-of Y', identifying this relation in text helps to design a first is-part-of extraction rule, which is used in turn to extracts new instances of that relation [29,40].

Two kinds of relations can roughly be distinguished: the generic ones, which can be found in almost any ontology, and the model-specific ones.

The links that form the main structure of an ontology are the most popular relations: the intra-concept relations (synonymy) and the hierarchical *is-a* and *part-of* relations. They can be considered either at the linguistic level (hyperonymy and meronymy are traditional lexicographic relations) or at the ontological level (is-a and part-of). The acquisition goal is to exploit the linguistic organization as it appears in texts to bootstrap the ontology design, even if the ontological structure is only partially reflected in the linguistic one. Various forms of extraction patterns have been designed to acquire such relations. See for instance the article of Cimiano *et al.* in this volume for examples of the application of such Hearst's patterns.

A wide range of domain specific relations are examined in IE works. Elementary relations can be interpreted as attributes of a given object class. The attributes age, name, phone number, parent, birthplace can be associated to a person [19]. Various relations can hold between objects or events: from semantic roles, such as agent or patient roles, to more complex ones such as the symptom relation in the medical domain or the interaction between biological entities in genomics.

Extracting relations between entities helps to populate a database. However, extracting a relation in isolation is usually not sufficient for ontology design. The elementary relation must be structured in more complex schemata [19,3]. For instance, in functional genomics, one of the most popular IE task aims at building enzymes and metabolic pathways, or regulation networks that can be considered as specific ontologies. Such networks are described by complex graphs of interactions between genes, proteins and environmental factors such as drugs or stress. The ontological result of the extraction should represent at least the entities, their reactions, their properties and, at a higher level, feedback cycles. Single elementary and binary relations between entities are independently extracted by IE methods. The integration of these elementary relations into the ontology highly depends on the biological model represented in an ontology and on the other extracted facts. Few works address this integration question. The improvement of an ontology by IE simply comes to add new instances of the interaction relation in most of the cases. For instance, with the semantic roles associated to *repress* (Agent(Repress, Protein) and Target(Repress, Gene)), the repress relation can be enriched by new instances. "SpoIIID represses spoVD transcription" yields Agent(Repress, SpoIIID) and Target(Repress, spoVD) [57]. Other works such as [46] aim at providing a user-friendly interface to facilitate the interpretation of the elementary results by the biologist.

4.2.1. Discussion

On the whole, although useful, pattern-based acquisition of relations cannot be the main knowledge source for ontology design. The best results in precision are obtained in hyponymy and specific relation extractions. Some reasons can be invoked. The variation in phrasing is difficult to capture and this affects the recall quality. General patterns must rely on grammatical words or construct (like prepositions) which are semantically vague. This affects the precision. More fundamentally, the linguistically based model cannot be directly mapped onto an ontology (see also [25]. Hyponymy between polysemous terms cannot be considered as a transitive relation; metonymy phenomena are conceptual shortcuts difficult to interpret; the language makes the confusion between the roles and the entities that hold the roles; etc. The use of relation extraction techniques must therefore be restricted to the complementation and tuning of an existing ontology and any extracted information needs to be further interpreted in ontological terms.

In the ExtraPloDocs project, we are currently investigating a method to combine the distributional analysis for learning synonymy and hyperonymy relations, which has a good coverage but produces noisy results with pattern-based relation extraction, which is more reliable but has a low productivity. As mentioned above, the distributional analysis is implemented in the Asium system, which produces a hierarchy of semantic classes of words. To improve the quality of the hierarchy produced by the Asium system and alleviate the validation burden, we aim at bootstrapping the distributional analysis with the various pieces of ontological knowledge which have been acquired by a pattern-based technique.

5. Conclusion

As illustrated above, the IE research related to ontologies is abundant, multiple and mainly applied. Many systems, approaches, algorithms and evaluations on quite basic applications are reported. At this stage, the main goal is more to develop systems that get a better precision and recall than making explicit and defending a given general approach against others. The influence of statistics on NLP, the influence of MUC on IE and the cost of ontological processing partially explain this. The simplest tasks are solved first (*e.g.* named entity recognition). IE methods for interpreting the lowest text levels are now well established. This maturity and the growing needs for real applications will draw the field towards a stronger involvement of the ontological knowledge.

Difficult and unexplored questions dealing with the discrepancy between what the text is about, the exogenous lexicon and a given ontology should be investigated. This gap may not be only due to representation languages, to divergent generality levels and incompleteness of the knowledge sources, which have been tackled by the revision field, but also to divergent text genres, points of view and underlying problem-solving tasks.

Ontology-driven IE and integration of the extracted knowledge in an ontology will not be properly done without appropriate answers to these questions.

References

[1] E. Alphonse and C. Rouveirol, Lazy propositionalisation for Relational Learning. *In 14th European Conference on Artificial Intelligence (ECAI'00, W. Horn ed.)*, Berlin, pp. 256-260, 2000.

[2] Alphonse E., Aubin S., Bessières P., Bisson G., Hamon T., Lagarrigue S., Nazarenko A, Manine A.-P., Nédellec C., Ould Abdel Vetah M., Poibeau T. and Weissenbacher D., Event-Based Information Extraction for the biomedical domain: the Caderige project". *Proceedings of the Workshop BioNLP (Biology and Natural language Processing)*, Conference on Computational Linguistics (Coling 2004), Geneva, pp. 43-49, 2004

[3] Aone C., Ramos-Santacruz M., REES: A Large-Scale Relation and Event Extraction System. *Proc. of ANLP'2000*, Seattle, 2000.

[4] Aussenac-Gilles N., Biébow B., Szulman S., Revisiting Ontology Design: a methodology based on corpus analysis. In R Dieng, O Corby (eds.) *Engineering and Knowledge Management: Methods, Models, and Tools. Proceedings of EKAW'2000*, LNAI 1937, Springer-Verlag, pp. 172-188, 2000.

[5] Basili R., Pazienza M.-T., Velardi P., Semi-automatic extraction of linguistic information for syntactic disambiguation, in *Applied Artificial Intelligence*, 7:339-364, 1993.

[6] Bessières P., Nazarenko N. and Nédellec C., Apport de l'apprentissageà l'extraction d'information : le problème de l'identification d'interactions géniques, in *Actes du Colloque International sur le Document Electronique, Méthodes, Démarches et Techniques Cognitives*, CIDE'2001, Toulouse, Octobre 2001.

[7] Bikel D. M., Miller S., Schwartz R., Weischedel R., Nymble: a High-Performance Learning Name-finder. *Conference on Applied Natural Language Processing*, 1997.

[8] Borthwick A., *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, Computer Science Department, New York University, 1999.

[9] Bourigault D., Gonzalez-Mulliez I., Gros C., LEXTER: A Natural Language Tool for Terminology Extraction. In *proceedings of the 7th Congress of EURALEX*, 1996.

[10] Chai Y. J., Biermann A. W., Guinn C. , Two dimensional generalization in IE. *Proceedings of the National Conference on Artificial Intelligence (AAAI'99)*, 1999.

[11] Chieu H. L., and Ng H. T., Named Entity Recognition: A Maximum Entropy Approach Using Global Information. *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*. pp. 190-196. Taiwan, 2002.

[12] Cimiano P., Ontology-driven discourse analysis in GenIE, in *Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems*, 2003.

[13] Cimiano P., Pivk A., Schmidt-Thieme L. and Staab S., Learning Taxonomic Relations from Heterogeneous Evidence. In *XXXXXXXTO BE COMPLETED*, IOS Press, 2005.

[14] Ciravegna F., Learning to Tag for Information Extraction from Text. *Proceedings of the ECAI-2000 Workshop on Machine Learning for Information Extraction*, F. Ciravegna et al. (eds), Berlin, 2000.

[15] Collier N., Ruch P. and Nazarenko A., *Proceedings of the Joint Coling workshop on Natural Language Processing in Biomedicine and its Applications*, 2004.

[16] Collier N., Nobata C., Tsujii J., Extracting the Names of Genes and Gene Products with a Hidden Markov Model. *Proceedings of COLING-2000*, Sarrebr§ck, 2000.

[17] Cowie J., Wilks Y., Information Extraction. In R. Dale, H. Moisl and H. Somers (eds.) *Handbook of Natural Language Processing*. New York: Marcel Dekker, 2000.

[18] *ExtraPloDocs Project*, French RNTL project funded by the Ministry of Research, http://www-lipn.univ-paris13.fr/ poibeau/ExtraPloDocs/. (2002-2005).

[19] Embley D. W., Campbell D. M., Smith R. D., Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Document. *Proceedings of CIKM'98*, 1998.

[20] Faure D. and Nedellec C., Knowledge acquisition of predicate argument structures from technical texts using Machine Learning: the system ASIUM. In *EKAW'99*, pp. 329-334, Springer-Verlag, 1999.

[21] Fillmore C. J., The case for case , in *Universals in linguistic theory*, Bachs & Harms (eds.), Holt, Rinehart and Winston, Chicago, 1968, pp 1-90.

[22] Freitag D., Toward General-Purpose Learning for Information Extraction. *Proceedings of COLING-ACL-98*, 1998.

[23] Fukuda K., Tamura A., Tsunoda T. and Takagi T., Toward information extraction: identifying protein names from biological papers. *Proceedings of PSB'98*. pp 707-18, 1998.

[24] Gaizauskas R., Wilks Y., Information Extraction: Beyond Document Retrieval. *Memoranda in Computer and Cognitive Science*, CS-97-10, 1997.

[25] Gangemi A., Guarino N., Oltramri A., Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level. *Proceedings of FOIS'2001*, Ogunguit, Maine, 2001.

[26] Grishman R. and Sterling J., Acquisition of Selectional Patterns, *Proceedings of the 14th. International Conference on Computational Linguistics* (COLING 92), 1992.

[27] Harris Z., *Mathematical Structures of Language*, John Wiley & Sons, New York, 1968

[28] Harris Z., Gottfried M., Ryckman T., Mattick P., Daladier A., Harris T. N. and Harris S. , *The Form of Information in Science: Analysis of an Immunology Sublanguage*, Kluwer Academic Publishers, Dordrecht, 1989.

[29] Hearst M. A., Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of COLING'92*, pp 539-545, 1992.

[30] Hobbs J. R., Appelt D., Bear J., Israel D., Kameyama M., Stickel M. and Tyson M., FASTUS: A Cascaded Finite-State Transducer for Extraction Information from Natural Language Text'. In E Roche and Y Schabes (eds.), *Finite-State Language Processing*, chapter 13, pp 383-406. MIT Press, 1997.

[31] Humphreys K., Demetriou G., Gaizauskas R., Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. In *Proceedings of PSB'2000*, 5:502-513, 2000.

[32] Huttunen S., Yangarber R. and Grishman R., Diversity of Scenarios in Information Extraction. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, 2002.

[33] Isozaki H. and Kazawa H., Efficient Support Vector Classifiers for Named Entity Recognition. *Proceedings of COLING-2002*, pp. 390-396, 2002.

[34] Jacquemin C., A Symbolic and Surgical Acquisition of Terms Through Variation. In S. Wermter, C. Bertrand and G. Scheler (eds), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Springer, pp. 425-438, 1996.

[35] Kim J., Moldovan D., Acquisition of linguistic patterns for knowledge-based information extraction. *IEEE Transacctions on Knowledge and Data Engineering*, 7(5):713-724, 1995.

[36] Leroy G. and Chen H., Filling preposition-based templates to capture information for medical abstracts. In *proceedings of PSB'2001*, Kaua'i, January, 2001.

[37] LLL'05 challenge: Genic Interaction Extraction Challenge, part of the Learning Language in Logic workshop, at ICML 2005. http://genome.jouy.inra.fr/texte/LLLchallenge/, Bonn, 2005.

[38] Meyer I., Skuce D., Bowker L., Eck K., Towards a new generation of terminological resources: an experiment in building a terminological knowledge base. *Proceedings of COLING'92*, Nantes, pp. 956-960, 1992.

[39] Miller G. A., WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235-312, 1990.

[40] Morin E. and Jacquemin C., Projecting Corpus-Based Semantic Links on a Thesaurus, *Proceedings of ACL'99*, Maryland. pp 389-396, 1999.

[41] Mikheev A., Feature Lattices for Maximum Entropy Modelling. In *proceedings of COLING-*

ACL, pp. 848-854, 1998.

[42] Mikheev A., Moens M. and Grover C., Named entity recognition without gazetteers. In *Proceedings of the Annual Meeting of the European Association of Computational Linguistics (EACL'99)*, Bergen, pp 1-8, 1999.

[43] MUC Proceedings, *Message Understanding conference*, 1987.

[44] Nazarenko A., Zweigenbaum P., Habert B. and Bouaud J., Corpus-based identification and refinement of semantic classes. *Journal of the American Medical Informatics Association* 4:584-589, 1997.

[45] Nédellec C., Machine Learning for Information Extraction in Genomics - State of the Art and Perspectives, *Text Mining and its Applications: Results of the NEMIS Launch Conference Series: Studies in Fuzziness and Soft Computing*, Sirmakessis, Spiros (Ed.), Springer Verlag, 2004.

[46] Ng S. and Wong M., Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*. 10:104-112, 1999.

[47] Ono T., Hishigaki H., Tanigami A. and Takagi T., Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*. 17(2): 155-161, 2001.

[48] Park J. C., Kim H. S. and Kim J. J., Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *proceedings of PSB'2001*, 2001.

[49] Proux D., Rechenmann F., Julliard L., Pillet V. and Jacq B., Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Informatics*. 9:72-80, 1998.

[50] Pustejovsky J., CastaŰo J., Zhang J., Kotecki M. and Cochran B., Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations. In *Proceedings of PSB'2002*, 7:362-373, 2002.

[51] Reinberger M.-L. and Spyns P., Discovering Knowledge in Text for the learning of DOGMA-inspired ontologies. In *XXXXXXXXTO BE COMPLETED*, IOS Press, 2005.

[52] Reyle U. and Saric J., Corpus Driven Information Extraction. In *Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Applications*, R. Baud and P. Ruch (eds), 2002.

[53] Riloff E., Automatically constructing a Dictionary for Information Extraction Tasks, *Proceedings of AAAI'93*, Washington DC, pp 811-816, 1993.

[54] Riloff E. and Sheperd K., A corpus-based approach for building semantic lexicons. In *Proceedings of EMNLP'97*, pp 117-124, 1997.

[55] Riloff E. and Jones R., Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, *Proceedings of AAAI-99*, pp 474-479, 1999.

[56] Rindflesch T. C., Tanabe L., Weinstein J. N. and Hunter L., EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. *Proceedings of PSB'2000*, vol 5:514-525, 2000.

[57] Roux C., Proux D., Rechenmann F. and Julliard L., An Ontology Enrichment Method for a Pragmatic Information Extraction System gathering Data on Genetic Interactions. *Proceedings of the ECAI'2000 Ontology Learning Workshop*, S Staab et al. (eds.), 2000.

[58] Ryu P.-M. and Choi K.-S., Measuring the Specificity of Terms for Automatic Hierarchy construction. In *XXXXXXXXTO BE COMPLETED*, IOS Press, 2005.

[59] Sager N., Friedman C. and Lyman M., *Medical Language Processing: Computer Management of Narrative Data*, Addison-Wesley, Reading, MA, 1987.

[60] Sasaki Y. and Matsuo Y., Learning Semantic-Level Information Extraction Rules by Type-Oriented ILP. *Proceedings of COLING-2000*, Kay M. (ed), Saarbr§cken, 2000.

[61] Schank R. and Abelson R., *Scripts, plans, goals and understanding*, Hillsdale, N.J. Lawrence Erlbaum, 1977.

[62] Sekimizu T., Park H. S. and Tsujii J., Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in MedLine Abstracts. In *Genome Informatics*.

Universal Academy Press Inc., Tokyo, Japan, 1998

[63] Soderland S., Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning Journal*, vol 34, 1999.

[64] Soderland S., Fisher D., Aseltine J. and Lehnert W., CRYSTAL: Inducing a Conceptual Dictionary. *Proceedings of IJCAI-95*, Montréal, pp 1314-1321, 1995.

[65] Takeuchi K. and Collier N., Use of Support Vector Machines in Extended Named Entity Recognition. *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan, August 2002.

[66] Thomas J., Milward D., Ouzounis C., Pulman S. and Carroll M., Automatic extraction of protein interactions from scientific abstracts. In *proceedings of PSB'2000*, pp 541-52, 2000.

[67] Uschold M., Knowledge modelling: concepts and terminology, in *The Knowledge Engineering Review*, 13(1), 5-29, 1998.

[68] Valencia A. and Blaschke C., *proceedings of the workshop "A critical assessment of text mining methods in molecular biology*, Spain, 2004.

[69] Yakushiji A., Tateisi Y., Miyao Y. and Tsujii .J-I., Extraction from biomedical papers using a full parser. *Proceedings of PSB'2001*, 2001.

[70] Yangarber R., Grishman R., Tapanainen P. and Huttunen S., Unsupervised Discovery of Scenario-Level Patterns for Information Extraction. *Proceedings of COLING-2000*, 2000.

[71] Zhou G. D. and Su J., Exploring Deep Knowledge Resources in Biomedical Name Recognition, in *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, 2004.