Intelligenza Artificiale 7 (2013) 125–137
DOI 10.3233/IA-130054
IOS Press

Semantic technologies for industry: From knowledge modeling and integration to intelligent applications

Giovanni Semeraro[a,*], Pierpaolo Basile[a], Roberto Basili[b], Marco de Gemmis[a], Chiara Ghidini[c], Maurizio Lenzerini[d], Pasquale Lops[a], Alessandro Moschitti[e], Cataldo Musto[a], Fedelucio Narducci[f], Arianna Pipitone[g], Roberto Pirrone [g], Piero Poccianti[h] and Luciano Serafini[c]
[a]*University of Bari Aldo Moro Bari, Italy*
[b]*University of Roma Tor Vergata, Rome, Italy*
[c]*Fondazione Bruno Kessler, Trento, Italy*
[d]*University of Roma La Sapienza, Rome, Italy*
[e]*Qatar Computing Research Institute and University of Trento, Trento, Italy*
[f]*University of Milan-Bicocca, Milan, Italy*
[g]*University of Palermo, Palermo, Italy*
[h]*Consorzio Operativo Gruppo Montepaschi, Sienna, Italy*

Abstract. Artificial Intelligence technologies are growingly used within several software systems ranging from Web services to mobile applications. It is by no doubt true that the more AI algorithms and methods are used the more they tend to depart from a pure "AI" spirit and end to refer to the sphere of standard software. In a sense, AI seems strongly connected with ideas, methods and tools that are not (yet) used by the general public. On the contrary, a more realistic view of it would be a rich and pervading set of successful paradigms and approaches. Industry is currently perceiving semantic technologies as a key contribution of AI to innovation. In this paper a survey of current industrial experiences is used to discuss different semantic technologies at work in heterogeneous areas, ranging from Web services to semantic search and recommender systems.The resulting picture confirms the vitality of the area and allows to sketch a general taxonomy of approaches, that is the main contribution of this paper.

Keywords: NLP, information retrieval, semantic search, recommender systems, ontologies, BPM

1. Introduction

Industries are currently facing with global markets while they are characterized by the design, production, synthesis and distribution of increasingly complex services and products. Governing such complexity is crucially tight to the availability of a variety of information, and competencies. Knowledge is here massively involved, and this strictly links with disciplines that have it at the center of their interest, such as Artificial Intelligence.

As supporters of big data analytics suggest, large scale empirical processes are certainly a promising direction to the development of knowledge in complex intelligent systems, but data are of no use without precise interpretation methods. Growing data volumes trigger increasingly complex interpretation problems

*Corresponding author: Giovanni Semeraro, University of Bari Aldo Moro, Bari, Italy. E-mail: semeraro@di.uniba.it.

over open domains. It does not come at a surprise to AI practitioners that knowledge is most often manifesting through unstructured forms, often dominated by uncertainty, vagueness and incompleteness. In this scenario, language is still the most powerful medium for knowledge acquisition, communication and sharing, optimized through centuries of successes and failures. It constitutes the preferred query language for naive users, early adopters and even domain experts.

Along the above view, the different contributions of this paper shed some light on the stringent interaction between knowledge and natural language. In the early sections, the role of natural language processing in enterprise technologies is widely discussed. Later sections go back to the role of structured forms of knowledge viewed as the elected sources in enterprise information system lifecycles. In particular, ontological data modeling and the semantic modeling of business processes are discussed.

2. Linguistic analysis and semantic resources in question answering systems

In this section important aspects of Natural Language Processing (NLP) for the design of competitive commercial Question Answering (QA) systems are highlighted. In particular, the role of advanced deep linguistic analysis and semantic resources as they become effective for Information Retrieval (IR) is discussed.

Until just about one and half decades ago, the impact of NLP and semantic resources in real-world applications was rather unsatisfactory. Many advanced text representations were proposed to improve indexing and retrieval of search engines, e.g., (i) *Noun Phrases* such as Named Entities (e.g., *George Bush* or *Washington D.C.*) and other complex nominals (e.g., *satellite cable television system*); (ii) tuples constituted by head words with their modifiers [67], where the relations between the tuple components are detected using syntactic parsers [16], e.g., *subject-verb*, *verb-object* and *subject-verb-object* (like in *Minister announces*, *announces plans* and *Minister announces plans*, respectively); and word senses, e.g., as organized in lexical resources like WordNet [23].

The aim of phrases was to improve the precision on concept matching, e.g., the index term constituted by the bigram <*company acquisition*> is more precise than the separated words <*company*> and <*acquisition*>. The aim of word senses was to improve on the precision of word matching, e.g., the verb *to raise* could refer to: (a) *agricultural texts*, when the sense is *to cultivate by growing* or (b) *economic activities* when the sense is *to raise costs*. Additionally, senses can be grouped together using synonyms or hypernyms by also improving on recall.

TREC conferences showed that phrases are ineffective for document retrieval, e.g., [66, 67]. The need of a Word Sense Disambiguation (WSD) algorithm for correctly using word senses was pointed out in [71] although promising results in IR were obtained with manual disambiguation [72]. However, it was shown that WSD was not enough accurate for improving retrieval [65]. On the text categorization side some improvement was derived in [8] when NLP was used in *weak* classifiers. However, simple bag-of-words models were shown to be in general more accurate [47] than NLP-derived representations. In summary, NLP for IR was fascinating but it could not be considered for real-world applications.

2.1. Effective NLP methods and resources for IR

Roughly seven-eight years ago, some signals on an imminent change in the perception of NLP were given by the growing exigency of providing more domain-specific information to the end user. For example, named entities started to play an important role in commercial applications, also thanks to potential business opportunities in the field of sentiment analysis. In particular, the latter was concretely shown to benefit from some basic syntactic processing [50]. Since then, the renewed interest in NLP has been growing along with changes in user needs, e.g., finer grain information extraction over large data sets rather than just document search.

The event that definitely assessed the importance of NLP for search and retrieval applications occurred in February 2011, when the IBM Watson system (hereafter referred as Watson), an advanced QA pipeline based on deep linguistic processing and semantic resources, demonstrated that automatic methods can be more accurate than human experts in searching and retrieving information. Additionally, the fast Watson's response made its search far more effective than the one operated by a human using automatic keyword-based methods.

It should be noticed that, in the NLP and IR perspective, the most important aspect is the essential role played by NLP for building Watson, rather than the victory over human champions. Such system could achieve an incredible accuracy thanks to the use of: (i) thousands of candidate answers made possible by

the extremely powerful computer clusters designed by IBM; (ii) advanced NLP and semantic resources [24]; and (iii) a reranker based on a machine learning algorithm for combining hundreds of NLP and semantic web techniques.

While (i) is interesting, it mainly regards software engineering optimization and distributed computation, which is beyond the scope of this section. (iii), surprisingly, just refers to typical logistic regressors applied to high-dimensional vectors. The second point is instead the *core* of the Watson technology.

In particular, several of the most effective features are built using two deep syntactic parsing components, an English Slot Grammar (ESG) parser and a predicate-argument structure (PAS) builder [45]. These are fundamental modules for question analysis, candidate generation, and analysis of passage evidence. They are also used for building additional core modules for relation extraction [56, 73]. In particular, syntactic information is used for identifying the type of the answer during question analysis; (ii) shallow semantic and syntactic information are used to give higher weight to terms connected to the question focus; and (iii) most importantly, the syntactic structure of the question is matched with the one of the answer passage to provide a compatibility feature. Semantic resources are also very important: the lexical answer type of the questions, e.g., as provided by Wikipedia, is matched against the one of the answer, whereas syntactic relations again play a major role for selecting answer and question keyword candidates.

Finally, PRISMATIC [22], a large-scale lexicalized relation resource, has had a major impact in providing features for answer scoring. It is interesting to notice that most important PRISMATIC frames are just triples of the form *subject-verb-object*, which were ineffective in early NLP-IR experiments. The reasons for their success in Watson are several: (i) their use is rather different than in [66] since they are features for supervised models, i.e., for the final reranker, applied to short texts. These latter result in much less noisy linguistic information than large document processing, which is exposed to many more errors. Additionally, supervised methods can filter noise whereas (unsupervised) document retrieval, cannot. (ii) Their quality is higher as they are obtained with better parsing technology by also using aggressive instance frequency filtering. (iii) They were extracted from effective resources, recently produced, e.g., Wikipedia, on very large scale (30 GBs of text). In [66], only a portion of TREC data could be used for efficiency reasons.

2.2. Beyond Watson

Watson has basically demonstrated that NLP is essential for high level IR tasks such as QA. This technology is also under consideration to improve traditional query search on the Web. It would seem that future work will address engineering rather than research aspects, therefore, this may raise the question: *Did the Watson performance de facto end basic research in QA?*

A straightforward answer is that there are interesting open problems concerning fast system prototyping and adaptation to different domains. Indeed, when passing from an application domain to another, NLP, especially if heavily based on manual feature or rule engineering is subject to large accuracy drops. Considering that syntactic representations based on feature vectors are difficult to design, especially when the answer is supported by multiple sentences, automatic methods for feature engineering are essential.

To reduce the burden of manual feature engineering for QA, we proposed structural models based on kernel methods, [48, 49, 56]. The main idea is to: (i) generate question and passage pairs, where the text passages are retrieved by a search engine; (ii) assume those containing the correct answer as positive instance pairs and all the others as negative ones; (iii) represent such pairs with syntactic/semantic trees; and (iv) apply learning to rank algorithms to sort answer passages by means of structural kernels. This approach enables the automatic engineering of structural/lexical semantic patterns. Finally, recent results [64] suggest that large-scale training is a promising direction to fast QA prototyping.

3. Semantic Search for Enterprise

Enterprise search has been defined as the process of searching within collections of digital textual materials owned by an organisation [33]. It includes search in company Intranets or specific websites.

Surveys suggest that small and medium size enterprises (SMEs) declare to make use of information when planning a technological innovation (e.g. [31] or [9]). The large majorty of SMEs seem to make use of Google, or general engines, by using keywords related to product types and functions, mostly through iterative searches. Timely and accurate access is crucial to innovation practices, so that organizations depend on robust autonomous filtering and classification capabil-

ities. Search here requires high level abstractions and a proactive role of the search system. Personalization is also found very important, as experts findings provide subjective behaviors, knowledge and preferences. Overall, one of the most requested features was related to the detection of patterns within documents able to expressing functional relationships and signaling innovative functions or customer requirements. In theories of process innovation, [1], these knowledge patterns are often called OAT tuples, i.e. Object-Action-Tool relations.

Semantic Enterprise Search involves, in this perspective, either ontological knowledge, to captures the company specific needs, as well as linguistic capabilities, aiming at characterizing the semantics of open text materials. Ontology-driven search applications are usually only moderately successful. Many applications emphasize the ontological structure to reformulate queries in query expansion, producing results comparable to thesaurus or dictionary-supported search solutions. Others encourage the use of semantic annotations to documents, but led to substantially more manual work. Semantic search is to be seen as a fusion of search engine with semantic web technologies, i.e. integrate semantic annotations (for intra-institutionwise distributed extensibility) while still maintaining free keyword search functionalities. A move toward more flexible semantic search is the idea of an IR engine with the capabilities to understand the user's intent and Web's content at a much deeper, conceptual level [46]. However, most ontology-driven search systems are tailored to some ontologies and are thus not application-independent.

There are two main directions that are relevant to the semantic search idea within organizations. The first is the approach to specialistic domains, where a bottom-up integration between lexical resources and knowledge bases is undertaken, and IR-like functionalities, such as query expansion or structured semantic similarity estimation for reranking is applied. It is the case of the INSEARCH [9]. The INSEARCH approach in the above sense stands in the integration of ontological knowledge (i.e. information expressed through the KR standards of RDF or OWL) with strongly lexicalized meaning representations, i.e. distributional models of the lexicons ([40, 69] or [7]). Vector models, widely used in Information Retrieval, are here applied to extend the lexical description of some concepts (such as SKOS-like topic categories or domain concepts), and drive statistical inferences during document classification or ranking. INSEARCH exploits the core framework of Semantic Turkey [57] a tool for semantic bookmarking/annotation, as a fully fledged Semantic Platform for Knowledge Management and Acquisition supporting all of W3C standards for Knowledge Representation (i.e. RDF/RDFS/OWL SKOS and SKOS-XL extension). Moreover, compositional distributional semantic models are used in INSEARCH to guide the user modeling of ontological concepts of interest (such as a SKOS topic), feed the document categorization process (that is sensitive to OAT patterns through vector based representation of their composition), concept spotting in texts as well as query completion. The adopted methods are discussed in [9], [3] and [7].

The second line looks at the integration of statistical language processing and ontology reasoning processes as it is often carried out for integration of structured and unstrctured data or QA against Open Linked Data. In general approaches, for QA against ontologies range between rule based (strongly deductive) systems to shallow approaches very close to the bag-of-words practices in document retrieval. In the first family of systems, we could mention at least PowerAqua [43] or Sindice [68]. Shortcomings of these approaches come from the naive (user's) dictionaries that can be very different from the data dictionary. In [70], a system that produces a SPARQL template to directly mirror the internal structure of the question and then instantiates the template using statistical entity identification and predicate detection is proposed. An interesting vocabulary independent approach is attempted in [26] which combines entity search, lexical similarity metrics to compute semantic relatedness and apply spreading activation onto RDF graphs. This work adopts a *lexicalist* perspective, which relies on a strong model of lexical semantic information to remove the ambiguity in the interpretation. Statistical inference combined with logic-based representation is also adopted in Semantic Parsing methods (e.g. [14]) where graphical models are used to select the correct interpretation in a predicate logic form. A similar path is followed by the University of Rome, Tor Vergata system, participating to the QALD-2013 competition, presented in [29]. This system interestingly combines probabilistic graphical models, lexical semantics and ontological inference within a robust QA architecture against Open Linked Data (in particular the DBpedia ontology targeted by the QALD competition). Statistical Inference is used to manage the ambiguity in the question. First, the *localization and retrieval of the ontological elements evoked by a question* is solved without relying on strict hypothesis on a static resource vocabulary. Distribu-

tional lexical semantics inference over words[1] links words of a question to concepts in the KB: ontological items are here *retrieved* in an IR-style, according to their semantic "*closeness*" to the question. Second, the system jointly solves the different ambiguities arising in the interpretation, by matching question grammatical structures and ontological information. In the interpretation process, an Hidden Markov Model is designed as a generative model of the question, i.e. *how a question is generated as a request against the RDF resource graph*. The decoding of the corresponding HMM jointly satisfies all constraints and select the suitable RDF graph. The resulting system does not rely on any ontology dependent resource with a clear beneficial impact on portability.

4. Cognitive Linguistics encounters semantic applications

Current standards for managing ontologies, such as OWL, are lacking in linguistic grounding, and are not able to achieve a clear link with natural language. Bridging this gap, unskilled users could be able to infer the information described in the ontology and it would be possible either producing or parsing utterances about the represented domain automatically. Moreover, as in the case of enterprises, it could be very useful not only accessing documents in the internal information system but also extracting information from external corpora that are related to the same domain.

Many attempts have been made with the aim of creating a *natural language interface* (NLI) to the ontology but very few of them use grammars; such interfaces are focused only on verbalizing information contained in the ontology, while it is often necessary to give exhaustive answers to the user queries by retrieving data outside of the knowledge base.

Existent approaches that query the ontology using natural language are based on controlled syntax; these approaches are an alternative to Controlled Natural Languages (CNL) [35, 61]. CNLs are subsets of natural languages, which are engineered in the sense that their grammar and vocabulary have been restricted systematically in order to reduce both ambiguity and complexity. Quelo [25] is an intelligent interface developed at the University of Bolzano for supporting the formulation of user requests. The authors define what a query is, how it is represented into the system, and the operations that are available to the user in order to modify the query. A feedback is provided, which presents only relevant information. The work is based on the idea presented in [21] where a technique is presented that allows users to access unstructured data sources by means of an integrated ontology. ORAKEL [15] is a NLI to knowledge base, which supports factual questions; these kinds of questions start with WH-pronouns such as "who", "what", "where", and so on. The term *factual questions* means that answers are ground facts as they are found in the knowledge base, and not complex answers to "why" or "how" questions, which require explanation. PANTO [75] is a portable NLI to ontologies. It accepts generic natural language queries and outputs SPARQL queries [59].

The NLI proposed by the research group at the University of Palermo relies on a quite different paradigm than the others mentioned above. In this case, the effort aims at avoiding both manual annotation and syntax limitation. The whole understanding/production process is conceived as the outcome of NL tasks performed by an artificial agent. In this perspective, the ontology has to be intended as the internal representation of the world (i.e. the domain) owned by the agent itself. Knowledge may increase over time as the result of several understanding tasks. The key idea underlying this approach is that the agent bounds its linguistic abilities for understanding and/or verbalizing about the domain to a core semantic resource, thus enriching its linguistic knowledge about the domain (the way of saying something).

The proposed system uses RDF/OWL ontologies to describe the domain, and a formalization of the RDF statements is presented, which describes ontological entities and their properties through a model inspired to Cognitive Linguistics [17, 41]. The Construction Grammar (CxG) theory [30, 38] has been used particularly. CxG considers together both semantics and syntax of either a grammatical or a lexical structure defining a "construction" as a form-meaning couple. Form and meaning are referred to as the *poles* of the construction.

The proposed approach infers *semantic seeds* from the RDF triples to build constructions, where a seed represents the meaning of the triple definition itself. If the triple defines a concept, an instance, or a relation, the semantic seed asserts the existence of such an entity in the domain through a first order logic predicate. A suitable set of rules based on linguistic typology has been designed to infer both semantics and syntax from the semantic seed. WordNet and FrameNet have been used as the reference linguistic resources. When one

[1] More on this topics in Section 5.

designs ontological resources like classes or properties, syntax is not controlled giving rise to multiple words labels without separators. An A* based algorithm is used to search the space of possible decompositions as a tree, and to extract relevant words from the semantic seed. Combining the inferred semantics and the syntax as the poles of constructions, a RDF triple becomes a grammatical construction or a lexical one depending on the semantic seed associated to it. The triple's semantics and its structure are the poles of such a construction. The whole set of constructions is made of the RDF triples in the ontology representation enriched with synonymic constructions derived from the base lexicon made by the RDF labels, and the grammatical constructions. It represent the lexical and grammatical knowledge of the system about the domain; such a knowledge enables the system to parse plain text and to produce utterances related to the domain.

The presented system computes all possible syntactic forms for the same meaning for building constructions. When retrieving information, such a behaviour allows semantic annotation of plain text as a side effect. In [58] the semantic annotator based on this methodology is shown; it is just one component of the full NLI system, and it can be used to expand the query results by obtaining data from external information sources.

The presented NLI has many potential applications besides annotation: it can return pieces of text from the processed documents as in the case of retrieving excerpts from administrative and legal corpora. Moreover, the system can be used to expand the ontology in use by adding new concepts and relations. Current research focuses on converting an entity-relation diagram (ERD) in OWL, thus allowing the system to verbalize about database contents and avoiding conventional SQL queries by human operators.

5. Semantics-aware Recommender Systems

Research on Semantics-aware Recommender Systems (SemRS) takes its rightful place at the intersection of AI, IR, NLP and Semantic Web (SW). Indeed, from an AI perspective, the behaviour of a Semantic Recommender System can be roughly cast as a ML problem whose goal is learning to categorize new items based on the above mentioned collection of observations. From an IR perspective, the information need, usually expressed by a query, is here represented by the user profile of the target user. In addition, most SemRSs deal with items described by features extracted from text, such as news, emails, or Web pages. Unlike structured data, there are no attributes with well-defined values. Text features involve a number of complications due to natural language ambiguity. In fact, keyword-based representations of items and user profiles, as well as string matching techniques to compare them, turn out to be insufficient to capture the semantics of user preferences and suffer of several problems, such as polysemy, synomymy and language dependence. These observations make very relevant the integration of proper techniques for deep content analytics borrowed from NLP and Semantic Technologies, which is one of the most innovative lines of research in SemRSs [37]. The key idea is the adoption of semantic indexing techniques that allow the shift from a keyword-based to a concept-based representation of items and user profiles. Semantic indexing techniques can be roughly divided into top-down and bottom-up approaches. The former rely on the integration of external knowledge, such as lexicons, folksonomies, sense inventories or ontologies, for performing word sense disambiguation [6], annotating items and representing user profiles in order to capture the semantics of the target user information needs. These techniques allow SemRSs to learn more accurate user profiles [63]. The latter exploit the so-called geometric metaphor of meaning to represent complex (syntagmatic and paradigmatic) relations between words in high-dimensional vector spaces [60]. The main motivation behind top-down approaches is the challenge of providing SemRSs with both the cultural background and linguistic knowledge, which characterizes the human ability of interpreting documents expressed in natural language and reasoning on their meaning (machine reading). A recent study [54] investigated the adoption of a technique named Explicit Semantic Analysis (ESA) [27] in SemRSs. The idea behind ESA is to view an encyclopedia as a collection of concepts (articles), each one provided with a large textual description (the article content). By ESA is thus possible to compute a semantic correlation value between a term (a word occurring in a Wikipedia article) and the Wikipedia articles themselves. The power of ESA is the capability of representing Wikipedia's knowledge base in a way that is directly used by a computer software, without the need for manually encoded common-sense knowledge. ESA was exploited for enriching document and profile representations of a SemRS by means of a feature generation process. Given a textual description (e.g., a movie plot), the feature generation process extracts the most related Wikipedia concepts (articles) for the input text. In this way, both

item descriptions and user profiles can be augmented with new concepts extracted from an encyclopedic source (i.e., Wikipedia). This process can lead to more transparent and serendipitous user profiles [55], and to more accurate item representations [54].

Techniques for deep content analytics have been also applied for knowledge infusion into recommender systems with the specific aim of overcoming the *overspecialization* or *serendipity problem* [36]. In fact, content-based algorithms provide suggestions of items already in the users' range of interests, limiting the discovery of new unknown and likely interesting items. In [62], the authors propose a strategy that allows to program for *serendipity*, i.e. that makes the system able to provide unexpected suggestions helping the user to find surprisingly interesting items which she might not have otherwise discovered. The recommendation algorithm is enhanced with a knowledge intensive process for providing the recommender system with the background knowledge useful for a deeper understanding of the items it deals with. The process extracts knowledge from different sources (Wikipedia, online dictionaries, WordNet) and creates a memory of linguistic and encyclopedic knowledge. A reasoning step based on spreading activation mechanisms [2] allows to obtain new knowledge in the form of words, which are then exploited by the recommendation algorithm to discover "hidden" associations between items, rather than computing a simple similarity score, and to produce non obvious suggestions.

Bottom-up approaches, that draw their inspiration from the geometric metaphor of meaning, rely on the so called distributional hypothesis [32], according to which the meaning of a word is determined by the rules of its usage in the context of ordinary and concrete language behaviour. This means that words are semantically similar to the extent that they share contexts. Hence, words are similar if they have the same contexts, that is to say, they are similar if they co-occur, where co-occurrence is defined with respect to a context, for example a document or an item description. This insight has been investigated to develop a content-based recommendation framework called enhanced Vector Space Model (eVSM) [51]. As in classical Vector Space Model, items and user profiles are represented as points in a vector space in eVSM as well. However, since VSM does not provide any semantic modeling of the information, distributional models were exploited to obtain a lightweight semantic representation of both items and user profiles, based on the co-occurences of the terms in the textual descriptions of the items. Furthermore, negative user preferences were represented by integrating the quantum negation operator proposed by Widdows, that took inspiration from logic connectives defined in Quantum Mechanics [76]. The effectiveness of the framework has been confirmed in several experimental settings, in both mono-lingual and multi-lingual evaluations [52]. Recently, the framework has been further extended to provide users with contextual recommendations as well, by exploiting the intuition that usage patterns of terms can be deeply analyzed to build a semantic vector space representation of the context itself, which can be combined with a classical representation of user preferences to generate context-aware recommendations.

6. Ontology-based data management

Figure 1 shows a portion of a relational table contained in a real, large information system. The table concerns the students of an Institution, where each row stores data about a single student. The first column contains her code (if the code is negative, then the record refers to a special student, called "fictitious"), columns 2 and 3 specify the time interval of validity for the record, ID_GROU indicates the group the student belongs to (if the value of FL_CP is "S", then the student is the leader of the group, and if FL_CF is "S", then the student is the controller of the group), AVERAGE is the average grade of the student (but the value is valid only if FLAG_AVG is "S"). Obviously, each notion mentioned above (like "fictitious", "group", "leader", etc.) has a specific meaning in the organization, and understanding such meaning is crucial if one wants to correctly manage the data in the table and extract information

ST_COD	TS_START	TS_END	ID_GROU	FL_CP	FLA_CF	AVERAGE	FLAG_AVG
4589	30/7/04	1/1/99	92736	S	N	25,45	N
0904	15/5/01	15/6/05	35060	N	N	23,06	N
4589	5/5/01	3/7/04	92736	N	S	19,50	S
-901	13/5/01	27/7/04	92770	S	N	29,20	N
9008	10/5/01	1/1/99	62010	N	S	24,70	S
-900	10/5/01	1/1/99	62010	S	N	20,00	N
0976	7/5/01	9/7/03	75680				

Fig. 1. A portion of a table in a database of a large organization.

out of it. Similar rules hold for the other 47 columns that, for lack of space, are not shown in the figure.

Those who have experience of large databases, or databases that are part of large information systems will not be surprised to see such complexity in a single data structure. Now, think of a database with many tables of this kind, and try to imagine a poor final user accessing such tables to extract useful information. The problem is even more severe if one considers that information systems in the real world use different (often many) heterogeneous data sources, both internal and external to the organization [11, 20].

6.1. Issues in governing complex information system

What the above example shows in simple form is that governing the resources (data, meta-data, services, processes, etc.) of modern information systems is still an outstanding problem. In particular, three important aspects related to this issue are discussed next.

Accessing and querying data. Although the initial design of a collection of data sources might be adequate, corrective maintenance actions tend to re-shape them into a form that often diverges from the original structure. The result is that the data stored in different sources and the processes operating over them tend to be redundant, mutually inconsistent, and obscure for large classes of users. So, query formulation often requires interacting with IT experts who knows where the data are and what they mean in the various contexts, and can therefore translate the information need expressed by the user into appropriate queries. On the other hand, it is often exceedingly difficult for end users to single out exactly the data that are relevant for them, even though they are perfectly able to describe their requirement in terms of business concepts.

Data quality. It is often claimed that data quality is one of the most important factors in delivering high value information services. However, the above-mentioned scenario poses several obstacles to the goal of even checking data quality, let alone achieving a good level of quality in information delivery.

Process and service specification. Information systems are crucial artifacts for running organizations, and organizations rely not only on data, but also, for instance, on processes and services. Designing, documenting, managing, and executing processes is an important aspect of information systems. However, specifying what a process/service does, or which characteristics it is supposed to have, cannot be done correctly and comprehensively without a clear specification of which data the process will access, and how it will possibly change such data.

6.2. OBDA: A new paradigm

In the last five years, several research groups, including the one at the University of Rome La Sapienza, have been working on a new paradigm addressing these issues, based on the use of knowledge representation and reasoning techniques. The paradigm [42] is called "Ontology-based Data Management" (OBDM), and requires structuring the information system into four layers.

- The *resource layer* is constituted by the existing data sources and applications that are relevant for the organization.
- *The knowledge layer* is constituted by a declarative and explicit representation of the whole domain of interest for the organization, called the domain knowledge base (DKB). The domain is specified by means of a formal and high level description of both its static and dynamic aspects, structured into four components: (*i*) the *ontology*, formally describing the information model of the organization and its basic usage primitives in terms of Description Logics [5], (*ii*) the specification of *atomic operations*, representing meaningful and relevant basic actions in the domain, (*iii*) the specification of *operating patterns*, describing the sequencing of atomic operations that are considered correct in the various contexts of the organization, and (*iv*) the *processes*, where each process is a structured collection of activities producing a specific service or product within the organization.
- The *mapping layer* is a set of declarative assertions specifying how the available resources map to the DKB.
- The *view layer* specifies views over the knowledge layer, both to be provided to internal applications, and to be exposed as open data and open APIs to third parties.

The distinguishing feature of the whole approach is that users of the system will be freed from all the details of how to use the resources, as they will express their needs in the terms of the DKB. The system will reason

about the DKB and the mappings, and will reformulate the needs in terms of appropriate calls to services provided by resources. Thus, for instance, a query will be formulated over the domain ontology, and the system will reason upon the ontology and the mappings to call suitable queries over data sources that will compute the answers to the original user query.

6.3. First experiences

A few research groups are experimenting OBDM in practice (see, for example, the Optique IP project, financed by the Seventh Framework Program (FP7) of the European Commission). The University of Roma La Sapienza is involved in applied projects both with Public Administrations, and with private companies. One of the experiences we are carrying out is with the Department of Treasury of the Italian Ministry of Economy and Finance [4]. In this project, three ontology experts from our department worked with three domain experts for six months, and built an ontology of 800 elements, with 3000 DL-Lite [12] axioms, and 800 mapping assertions to about 80 relational tables. The ontology is now used as a common framework for all the applications, and will constitute the main document specifying the requirement for the restructuring of the information system that will be carried out in the next future.

7. Semantic Business Process Modeling

Semantic Business Process Management [34] aims at improving the level of automation in the specification, implementation, execution, and monitoring of business processes by extending business process management tools with the most significant results from the area of Semantic Web. When the focus is on process modelling, i.e., the activity of specification of business processes at an abstract level (descriptive and non executable), annotating process descriptions with labels taken from a set of domain ontologies, or enriching the process description with data objects taken from a set of domain ontologies, provides additional support to business experts in their modeling activities, including the modeling of valid diagrams which satisfy semantically enriched and domain specific constraints. A clear demonstration of this, is the stream of recent work on the introduction and usage of formal semantics to support Business Process Management [10, 18, 19, 39, 77].

Analyzing this stream of work we can roughly divide the different approaches into two groups: (i) those adding semantics to specify the dynamic behavior exhibited by a business process [39, 75, 77], and (ii) those adding semantics to specify the meaning of the entities of a business process in order to improve the automation of business process management [10, 18, 19]. The approach we take in this section, which briefly summarizes the work introduced in [28] and applied in [13], belongs to the second group and provides an example of usage of semantic web technology, and in particular Description Logics (DLs) [5], to specify and verify *structural* requirements, that is, requirements which refer to *descriptive* properties of the annotated process diagram and not to its execution.

7.1. Representing Semantically Annotated Processes

Consider the starting portion of a process for applying for care financial support in an Italian health service organization depicted in the Business Process Diagram (BPD) of Fig. 2 composed using the BPMN notation. This BPD focuses on the interaction between the applicant and the fiscal office in charge of checking whether the applicant is in a status of economic need, and if so to send the request to the health service for the medical checks.

The modelling of this process may involve business designers and analysts who may wish to impose and verify requirements on the process itself. The requirements, or constraints, may refer to the actions performed in the process (e.g., *"the evaluation of a certified income must follow the production of an income certification"*) as well as to the actors performing certain actions, or to the documents (data structures) manipulated in the process (e.g., *"only CAAFS can produce income certifications"*). Specifying and verifying these requirements demands for the ability to "understand" the domain specific semantics of the elements used in the BPD, that is, for instance the fact that certain actions are *"evaluations of certificates"*, that certain data objects are *"ICEF certificates"* and so on.

The work presented in [28] provides both: (i) a framework for the representation of semantically annotated business processes and of structural constraints by means of DL ontologies, and (ii) a modeling tool to support the collaborative creation of semantically annotated BPDs.

The formal representation is achieved by representing semantically annotated BPDs and structural constraints by means of the *Business Process Knowledge Base* (BPKB), schematised in Figure 3. Roughly speaking, the T-box contains a formalisation of BPMN

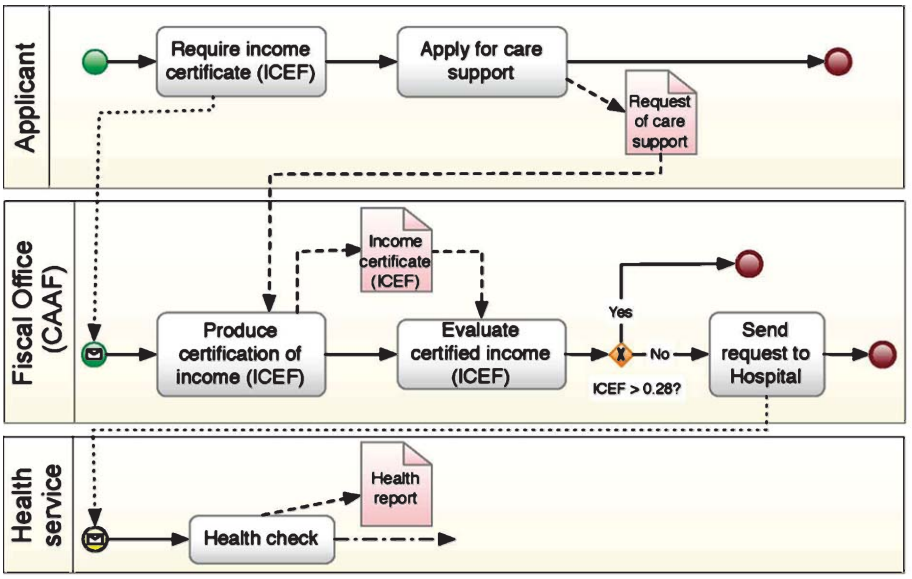

Fig. 2. A sample process.

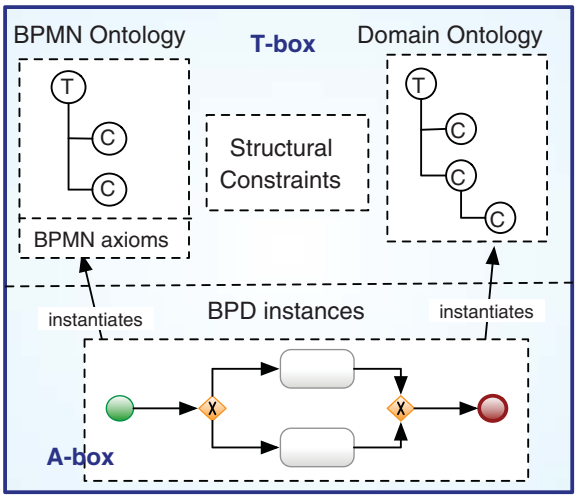

Fig. 3. The BPKB.

(*BPMN ontology*), a specific business *domain ontology* used to annotate the PBD, and a specification of the constraints to be satisfied, while the A-box contains the the actual elements of the annotated BPD in terms of instances of the BPMN and the domain ontology. The reasoning tasks are performed with the support of the Pellet reasoner, integrated with the Pellet IC Validator for the constraint validation tasks.

The modeling tool, called MoKi[2], enables the modeling of ontological and procedural knowledge in a collaborative and integrated manner using a set of Wiki pages (based on MediaWiki [44]). The tool offers graphical facilities for process definition and for ontology visualization and editing, as well as an automatic export of the graphical annotated BPDs as instances of the BPKB illustrated above.

7.2. Semantic Business Process Modeling in Action

In the last few years, the Public Administrations (PA) of several countries have invested effort and resources into modernizing their services, for instance by replacing their paper-based documents with electronic-based ones. The availability of precise models of the procedures of the PA and of specific "entities" related to these procedures, such as the *documents* produced by the procedures or the *organizational roles* performing the activities, is a a key factor towards both (1) the re-design of the administrative procedures in order to replace paper-based documents with electronic-based ones, and (2) the definition of guidelines and functions needed to safely store, catalogue, manage and retrieve the electronic documents in an appropriate archival systems.

In [13] we report the experience of using MoKi in the context of the ProDe Italian national project[3] to involve domain experts in producing precise models precise models of the procedures of the PA by means of semantically annotated BPDs.

[2] http://moki.fbk.eu

[3] http://www.progettoprode.it/Home.aspx

The quantitative data collected on the *usage* and *evaluation* of MoKi within the Prode project highlight the potential and criticality of using semantic wiki-based tools for building semantically annotated processes in real settings. In brief we can state that the users perceive the tool as more than easy to use. This result is also strengthened by the fact that 72% of employees spent only less than two days to learn how to use MoKi, and the same percentage learned it autonomously. Moreover, we observed that users positively perceive the overall usefulness of the tool for the collaborative modeling of documents and processes. The validity of this result is also confirmed by the fact that such a usefulness is perceived more strongly by employees working in teams having more than two persons. There exists, in fact, a correlation between the size of the subject's team and his/her feedback about the MoKi usefulness for collaborative purposes.

Acknowledgements

We would like to thank the researchers of the IBM Watson Research Center, Yorktown Heights, NY for the innumerable nice chats on Watson and NLP topics. Part of the author's research outlined in Section 2 has been supported by an Open Collaborative Research (OCR) award from IBM Research.

References

[1] G. Altshuller, *40 principles, TRIZ keys to technical innovation*. Number 1 in Triz tools. Technical Innovation Center, Worcester, Mass., 1. ed edition, 1998.

[2] J.R. Anderson, A spreading activation theory of memory, *Journal of Verbal Learning and Verbal Behavior* **22**:261–295, 1983.

[3] P. Annesi, V. Storch and R. Basili, Space projections as distributional models for semantic composition, In *CICLing (1)*, 323–335, 2012.

[4] N. Antonioli, F. Castanò, C. Civili, S. Coletta, S. Grossi, D. Lembo, M. Lenzerini, A. Poggi, D. Savo and E. Virardi, Ontology-based data access: The experience at the italian department of treasury, In *CAISE*, 2013.

[5] F. Baader, D. Calvanese, D. McGuinness, D. Nardi and P.F. Patel-Schneider, editors. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, 2003.

[6] P. Basile, G. Semeraro, M. Degemmis, A.L. Gentile, P. Lops and G. Semeraro, UNIBA: JIGSAW algorithm for word sense disambiguation, In *SemEval-2007*, pp. 398–401, 2007.

[7] R. Basili, C. Giannone and D. De Cao, Learning domainspecific framenets from texts, In *ECAI Workshop on Ontology Learning and Population*, 2008.

[8] R. Basili, A. Moschitti and M. Pazienza, NLP-driven IR: Evaluating performances over text classification task, In *IJCAI01*, 2001.

[9] R. Basili, A. Stellato, D. Previtali, P. Salvatore and J. Wurzer, Innovation-related enterprise semantic search: The INSEARCH experience, In *ICSC*, pp. 194–201, 2012.

[10] C. Beeri, A. Eyal, S. Kamenkovich and T. Milo, Querying business processes, In *VLDB 2006*, 343–354, 2006.

[11] P.A. Bernstein and L. Haas, Information integration in the enterprise, *Communications of the ACM* **51**(9):72–79, 2008.

[12] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini and R. Rosati, Tractable reasoning and efficient query answering in description logics: The DL-Lite family, *Journal of Automated Reasoning* **39**(3):385–429, 2007.

[13] C. Casagni, C. Di Francescomarino, M. Dragoni, L. Fiorentini, L. Franci, M. Gerosa, C. Ghidini, F. Rizzoli, M. Rospocher, A. Rovella, L. Serafini, S. Sparaco and A. Tabarroni, Wikibased conceptual modeling: An experience with the public administration. In *ISWC 2011*, pp. 17–32, 2011.

[14] D.L. Chen and R.J. Mooney, Learning to interpret natural language navigation instructions from observations, In *AAAI 2011*, pp. 859–865, 2011.

[15] P. Cimiano, P. Haase and J. Heizmann, ORAKEL: A portable natural language interface to knowledge bases. Technical report, Institute AIFB, University of Karlsruhe, 2007.

[16] M. Collins, Three generative, lexicalized models for statistical parsing, In *Proceedings ACL and EACL*, pp. 16–23, 1997.

[17] W. Croft, *Syntactic categories and grammatical relations: The cognitive organization of information*, University of Chicago Press, 1991.

[18] A. De Nicola, M. Lezoche and M. Missikoff, An ontological approach to business process modeling, In *IICAI 2007*, pp. 1794–1813, 2007.

[19] M. Dimitrov, A. Simov, S. Stein and M. Konstantinov, A BPMO Based Semantic Business Process Modelling Environment, In *SBPM 2007*, volume 251 of CEUR-WS, 2007.

[20] A. Doan, A.Y. Halevy and Z.G. Ives, *Principles of Data Integration*, Morgan Kaufmann, 2012.

[21] P. Dongilli and E. Franconi, An intelligent query interface with natural language support, In *FLAIRS*, pp. 658–663, AAAI Press, 2006.

[22] J. Fan, D. Ferrucci, D. Gondek and A. Kalyanpur, Prismatic: Inducing knowledge from a large scale lexicalized relation resource. In *NAACL HLT 2010*, pp. 122–127, 2010.

[23] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[24] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J.W. Murdock, E. Nyberg, J. Prager, N. Schlaefer and C. Welty, Building Watson: An Overview of the DeepQA Project, *AI Magazine*, **31**(3):59, 2010.

[25] E. Franconi, P. Guagliardo, M. Trevisan and S. Tessaris. Quelo: An ontology-driven query interface, In *Description Logics*, 2011.

[26] A. Freitas, J.G. Oliveira, S. O'Riain, E. Curry and J.C.P. Da Silva. Querying linked data using semantic relatedness: A vocabulary independent approach, In *NLDB 2011*, pp. 40–51, 2011.

[27] E. Gabrilovich and S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, *Journal of Artificial Intelligence Research* **34**:443–498, 2009.

[28] C. Ghidini, C.D. Francescomarino, M. Rospocher, P. Tonella and L. Serafini, Semantics based aspect oriented management of exceptional flows in business processes, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **42**(1):25–37, January 2012.

[29] C. Giannone, V. Bellomaria and R. Basili, A HMM-based Approach to Question Answering against Linked Data. In *Working Notes of the QALD-3 Challenge at CLEF*, 2013.
[30] A.E. Goldberg, *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.
[31] G. Grefenstette, Upcoming industrial needs for search. In *ECIR*, p. 3, 2009.
[32] Z.S. Harris, *Mathematical Structures of Language*. Interscience, New York, 1968.
[33] D. Hawking, Challenges in enterprise search. In *ADC '04*, pp. 15–24, 2004.
[34] M. Hepp, F. Leymann, J. Domingue, A. Wahler and D. Fensel, Semantic business process management: A vision towards using semantic web services for business process management. In *ICEBE 2005* IEEE, pp. 535–540, 2005.
[35] W.-O. Huijsen, Controlled language - an introduction. In *Proceedings of CLAW*, 1998.
[36] L. Iaquinta, M. Degemmis, P. Lops, G. Semeraro, M. Filannino and P. Molino. Introducing serendipity in a content-based recommender system. In *HIS 08* IEEE, pp. 168–173, 2008.
[37] D. Jannach, M. Zanker, A. Felfernig and G. Friedrich, *Recommender systems: An introduction*, Cambridge University Press, 2010.
[38] P. Kay and C.J. Fillmore, Grammatical constructions and linguistic generalizations: The What's X Doing Y? construction, *Language* **75**(1):1–33, 1999.
[39] A. Koschmider and A. Oberweis, Ontology based business process description. In *CAiSE-05 Workshops, LNCS*, pp. 321–333. Springer, 2005.
[40] T.K. Landauer and S.T. Dumais, A Solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**(2):211–240, 1997.
[41] R.W. Langacker, *Concept, Image, and Symbol: The Cognitive Basis of Grammar, volume 1 of Cognitive Linguistic Research*. Second edition, 2002.
[42] M. Lenzerini, Ontology-based data management. In *CIKM ACM*, 2011.
[43] V. Lopez, A. Nikolov, M. Sabou, V.S. Uren, E. Motta and M. d'Aquin. Scaling up question-answering to linked data. In *EKAW*, pp. 193–210, 2010.
[44] Wikimedia Foundation. Mediawiki. http://www.mediawiki.org, Accessed on 6 Nov 2011.
[45] M.C. McCord, J.W. Murdock and B. Boguraev, Deep parsing in Watson, *IBM Journal*, **56**(3):264–278, 2012.
[46] P. Mika, Microsearch: An interface for semantic search. In *SemSearch*, pp. 79–88, 2008.
[47] A. Moschitti, A study on convolution kernels for shallow semantic parsing. In *ACL*, pp. 335–342, 2004.
[48] A. Moschitti, Kernel methods, syntax and semantics for relational text categorization, In *CIKM*, 2008.
[49] A. Moschitti, J. Chu-Carroll, S. Patwardhan, J. Fan and G. Riccardi. Using Syntactic and Semantic Structural Kernels for Classifying Definition Questions in Jeopardy! In *EMNLP01*, pp. 712–724.
[50] T. Mullen and N. Collier, Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, pp. 412–418, 2004.
[51] C. Musto, Enhanced vector space models for content-based recommender systems. In *RecSys 2010*, pp. 361–364, ACM, 2010.
[52] C. Musto, F. Narducci, P. Basile, P. Lops, M. de Gemmis and G. Semeraro, Cross-language information filtering: Word sense disambiguation vs. distributional models. In *AI*IA 2011*, volume 6934 of *LNCS*, pp. 250–261, Springer, 2011.
[53] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J.H.M. Korst, V. Pronk and R. Clout, Enhanced semantic tv-show representation for personalized electronic program guides. In *UMAP 2012*, volume 7379 of *LNCS*, pp. 188–199, Springer, 2012.
[54] F. Narducci, *Knowledge-enriched Representations for Content-based Recommender Systems*, PhD thesis, University of Bari Aldo Moro, 2012.
[55] F. Narducci, C. Musto, G. Semeraro, P. Lops and M. de Gemmis. Leveraging encyclopedic knowledge for transparent and serendipitous user profiles. In *UMAP 2013*, volume 7899 of *LNCS*, pp. 350–352, Springer, 2013.
[56] T.-V.T. Nguyen and A. Moschitti, Joint distant and direct supervision for relation extraction, In *IJCNLP*, pp. 732–740, 2011.
[57] M. T. Pazienza, N. Scarpato, A. Stellato and A. Turbati, Semantic turkey: A browser-integrated environment for knowledge acquisition and management, *Semantic Web Journal*, **3**(3):279–292, 2012.
[58] A. Pipitone and R. Pirrone, Cognitive linguistics as the underlying framework for semantic annotation. In *ICSC*, IEEE Computer Society, pp. 52–59, 2012.
[59] E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF. Technical report, 2008.
[60] M. Sahlgren, *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University, Department of Linguistics, 2006.
[61] R. Schwitter and M. Tilbrook, Meaningful web annotations for humans and machines using controlled natural language, *Expert Systems*, **25**(3):253–267, 2008.
[62] G. Semeraro, P. Lops, P. Basile and M. de Gemmis, Knowledge infusion into content-based recommender systems. In *RecSys 2009*, pp. 301–304, ACM, 2009.
[63] G. Semeraro, P. Lops and M. Degemmis, WordNet-based user profiles for neighborhood formation in hybrid recommender systems. In *HIS 05*, pp. 291–296. IEEE, 2005.
[64] A. Severyn and A. Moschitti, Structural relationships for largescale learning of answer re-ranking, In *SIGIR*, 2012.
[65] A.F. Smeaton, Using NLP or NLP resources for information retrieval tasks. In *Natural language information retrieval*, pp. 99–111, Kluwer Academic Publishers,1999.
[66] T. Strzalkowski, J.P. Carballo, J. Karlgren, A. Hulth, P. Tapanainen and T. Jarvinen, Natural language information retrieval: TREC-8 report, In *TREC*, 1999.
[67] T. Strzalkowski and S. Jones, NLP track at TREC-5. In *Text REtrieval Conference*, 1996.
[68] G. Tummarello, R. Delbru and E. Oren, Sindice.com: Weaving the open linked data. In *ISWC*, pp. 552–565, 2007.
[69] P.D. Turney and P. Pantel, From frequency to meaning: Vector space models of semantics, *Journal of Artificial Intelligence Research*, **37**:141–188, 2010.
[70] C. Unger, L. Bühmann, J. Lehmann, A.-C.N. Ngomo, D. Gerber and P. Cimiano, Template-based question answering over rdf data, In *WWW*, pp. 639–648, 2012.
[71] E. M. Voorhees, Query expansion using lexical-semantic relations, In *SIGIR*, 1994.
[72] E.M. Voorhees, Using wordnet for text retrieval. In C. Fellbaum, editor, *Word Net: An Electronic Lexical Database*, The MIT Press, pp. 285–303, 1998.
[73] C. Wang, J. Fan, A. Kalyanpur and D. Gondek. Relation extraction with relation topics, In *EMNLP*, pp. 1426–1436, 2011.

[74] C. Wang, M. Xiong, Q. Zhou and Y. Yu, PANTO: A portable natural language interface to ontologies, In *ESWC 2007*, pp. 473–487, 2007.
[75] I. Weber, J. Hoffmann and J. Mendling, Semantic business process validation. In *SBPM 2008*, 2008.
[76] D. Widdows, Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *ACL 03*, pp. 136–143, 2003.
[77] P. Wong and J. Gibbons, A Relative Timed Semantics for BPMN. In *FOCLASA*, 2008.