SOMMARIO

L'obiettivo principale di questa ricerca è proporre un approccio strutturalista per l'elaborazione automatica della conoscenza attraverso l'apprendimento e il popolamento di ontologie, realizzata da/per testi strutturati e non strutturati. Il metodo suggerito include approcci di semantica distribuzionale e teorie di formalizzazione dei linguaggi naturali, al fine di sviluppare un quadro di riferimento che si basa su un'analisi linguistica *fine-grained*.

Partendo da una panoramica degli algoritmi di apprendimento automatico più diffusi e degli approcci basati su regole, presenteremo una metodologia per la creazione di un parallelismo tra formalismi macchina e modelli linguistici. In particolare, nella sezione 1, faremo una breve introduzione su alcuni concetti fondamentali, come la conoscenza, la rappresentazione e il ragionamento logico. Successivamente, si prenderà in considerazione la relazione esistente tra rappresentazioni formali e i linguaggi naturali e si introdurranno le norme per gli schemi di metadati e i modelli concettuali disponibili per il dominio dei Beni Culturali (BBCC).

Nella sezione 2, per affrontare i principali compiti relativi all'elaborazione automatica della conoscenza basata sulle ontologie, useremo la definizione di ontologia, richiamando anche la sua struttura e gli obiettivi.

Nella sezione 3, introdurremo alcuni dei principali metodi stocastici/statistici utilizzati per l'estrazione della conoscenza e dell'informazione attraverso ontologie. Per ciascuna delle tecniche presentate, forniremo una descrizione accurata, insieme ad alcuni esempi di applicazioni specifiche.

Nella sezione 4, in riferimento all'apprendimento e al popolamento di ontologie, introdurremo i principali modelli e i metodi utilizzati in compiti di trattamento automatico del linguaggio naturale che si basano su diversi tipi di *framework*. Infatti, al fine di analizzare le lingue naturali e storiche, l'elaborazione linguistica guida il livello di analisi, che può riguardare - contemporaneamente o separatamente - i tre diversi strati pertinenti a fonologia, sintassi e semantica.

Per quanto riguarda questi argomenti, al punto 5, proporremo il nostro approccio, basato sul quadro teorico del Lessico-Grammatica (LG), per il raggiungimento della formalizzazione del linguaggio naturale nel dominio di conoscenza dell'Archeologia. Intendiamo dimostrare come la nostra tecnica di formalizzazione linguistica può essere applicato sia al processo che al popolamento di un'ontologia di dominio, che mira a sviluppare un trattamento della conoscenza efficiente. La nostra formalizzazione linguistica si basa su un'osservazione accurata delle proprietà lessicali, e su un'appropriata registrazione dei dati linguistici di tutto il lessico e dei comportamenti combinatori delle entrate lessicali, includendo la sintassi e anche il lessico. Si differenzia dalle più conosciute tra le teorie linguistiche, come per esempio la grammatica profonda di Chomsky e le sue diverse derivazioni, che sono fortemente formali e basate sulla sintassi. Al fine di creare le principali risorse linguistiche da applicare nel nostro sistema, durante l'elaborazione linguistica, è stato sviluppato l'*Archaeological Italian Electronic Dictionary* (AIED). Inoltre, sono state create altre risorse linguistiche adatte ad applicare i vincoli semantici e ontologici che guidano le analisi linguistiche e i processi di estrazione.

Nella sezione 6, presenteremo il workflow di sistema che intendiamo sviluppare al fine di integrare le nostre risorse linguistiche in un ambiente adatto per un motore di ricerca semantico, chiamato Endpoint for Semantic Knowledge (ESK). ESK è strutturato come un endpoint SPARQL, che applica una analisi semantica *fine-grained*, basata sullo sviluppo di un modello di correlazione tra una serie di formalismi semantici per le macchine e un insieme di frasi in linguaggio naturale. ESK consente agli utenti di interrogare in linguaggio naturale una base di conoscenza, come DBpedia e Europeana, e di elaborare testi non strutturati, sia caricati dagli utenti che acquisiti on line, al fine di rappresentare e estrarre conoscenza.

Infine, chiuderemo la nostra ricerca valutando i suoi risultati e presentando possibili prospettive di lavoro future.

Keywords:
Elaborazione Conoscenza, TAL, Popolamento Ontologie, Apprendimento Ontologie, Modelli Linguistici Formali, Lessico-Grammatica.

ABSTRACT

The main aim of this research is to propose a structuralist approach for knowledge processing by means of ontology learning and population, achieved starting from unstructured and structured texts. The method suggested includes distributional semantic approaches and NL formalization theories, in order to develop a framework, which relies upon deep linguistic analysis.

Starting from an overview of the most spread machine learning algorithms and rule-based approaches, we will present a methodology for creating a parallelism between machine formalisms and linguistic models.

More specifically, in section 1, we will make a brief introduction to some core concepts, such as knowledge, representation and logic reasoning. Subsequently, we will consider the relationship between formal representations and natural languages and we will introduce standards for metadata schemata and conceptual models available for the Cultural Heritage (CH) domain.

In section 2, to deal with the main tasks related to ontological Knowledge Processing (KP), we will use the definition of ontology, also recalling its structure and goals.

In section 3, we introduce some of the main stochastic/statistical methods used to extract knowledge and information through ontologies. For each of the technique presented, we will provide an accurate description, together with some samples of specific applications.

In section 4, as for ontology learning and population, we will introduce the main models and methods used in NLP tasks and which are based on different types of frameworks. Actually, in order to analyse natural and historical tongues, linguistic processing addresses the level of the analyses, which may concern – contemporarily or separately – the three different layers of phonology, syntax and semantics.

As for these topics, in section 5, we will propose our approach, based on Lexicon-Grammar (LG) framework, to the achievement of natural language formalizations in the Archaeological knowledge domain. We intend to demonstrate how our language formalization technique can be applied to both process and populate a domain ontology, aiming at developing an efficient and effective knowledge processing. Our linguistic formalization is based on an accurate observation of lexical properties, and

on an appropriate linguistic data recording of all lexicon and lexical entry combinatory behaviours, encompassing syntax and, also, lexicon. It differs from the best known among current linguistic theories, as for instance Chomsky's deep grammar and its various offspring, which are strictly formalist and syntax-based.

The Archaeological Italian Electronic Dictionary (AIED) has been developed in order to create the main Linguistic Resources which are applied in our system during linguistic processing.

Furthermore, we create other resources suitable to the application of semantic and ontological constraints which drive linguistic analyses and extraction processes.

In section 6, we will present the system workflow we intend to develop in order to integrate our LRs in an environment suitable for a semantic search engine, called Endpoint for Semantic Knowledge (ESK). ESK is structured as a SPARQL endpoint, which will be applying a deep semantic analysis, based on the development of a matching model between a set of machine semantic formalisms and a set of NL sentences.

ESK allows users to run an NL query against KBs, such as DBpedia and Europeana, and to process unstructured texts, both uploaded by users and retrieved on line, in order to represent and extract knowledge.

Finally, we will close our research evaluating its results and presenting possible future work perspectives.

Keywords:

Knowledge Processing, Natural Language Processing, Ontology Population, Ontology Learning, Linguistic Formal Models, Lexicon-Grammar.