A Semantic Annotation Model for Syntactic Parsing in Ontology Learning Task

Maria Pia di Buono, Mario Monteleone
DISPSC
UNISA
Fisciano (SA), Italy
{mdibuono,mmonteleone}unisa.it

***Abstract*—In this paper, we present an on-going research on a semantic annotation model, which aims at creating a system for syntactic parsing apt to achieve Ontology Learning (OL) tasks. Our approach is based on a deep linguistic analysis, derived from predicate-argument structures and established on the notion *de prédicat sèmantique* (semantic predicate), following Maurice Gross' approach.**

Keywords - Ontology Learning; Deep Linguistic Analysis; Syntactic Parsing; Lexicon-Grammar.

I. INTRODUCTION

Semantic annotation process of sentence structures is one of the most challenging Natural Language Processing (NLP) tasks. Usually, applied techniques may be distinguished by the use of shallow or deep semantic analysis [1]. Besides, some recent approaches apply hybrid methods [1], in order to exploit the benefits derived from both for each step of the process. Shallow linguistic processing concerns the achievement of specific NLP tasks, even if it does not aim at accomplishing an exhaustive linguistic analysis. Systems based on a shallow approach are generally oriented to tokenization, part-of-speech tagging, chunking, named entity recognition, and shallow sentence parsing. Due to the improvement of such systems, in the last years the capability of text analysis achieved by shallow techniques has increased. Still, in terms of efficiency and robustness, shallow technique results are not comparable to deep system ones.

On the other hand, deep linguistic processing mainly refers to approaches, which apply linguistic knowledge to analyze natural languages. Such linguistic knowledge is encoded in a declarative way, namely in formal grammars, yet neither in algorithms nor in simple database. Thus, a formal grammar becomes an expression of both a certain linguistic theory and some operations, which are used to check consistence and to define information fusing. For this reason, deep linguistic processing is usually defined as a rule-based approach. Due to the fact that statistical methods may also be applied to deep grammars and systems, this does not mean that rule-based approaches are opposite to statistical methods. In deep linguistic processing, rules state constraints, based on a linguistic theory, which drives correct syntax of linguistic entities, while words are encoded in a specific lexicon. Syntax rules are not only related to grammatical correctness, on the basis of which a sentence is grammatically approved or rejected , but they may also describe semantic representations. Thus, syntax seems to be able to express both linguistic levels, namely the grammatical level and the meaning one.

The rest of the paper is structured as follows. In Section II, we present the most widely used annotated corpora, the Penn TreeBank and the PropBank. Then, in Section III, we introduce the main goals of OL, highlighting the challenging aspects in the achievement of such task. Consequently, in Section IV, we propose our framework to develop a system for syntactic parsing. Finally, in Section V, we analyze the proposed annotation process, which is based on a semantic tagging.

II. RELATED WORKS

The most well-known annotated corpora are the Penn TreeBank [2] and the PropBank [3].

The Penn TreeBank (1989-1996), is a parsed corpus, syntactically and semantically annotated, which produces:

- 7 million words of part-of-speech tagged text,
- 3 million words of skeletally parsed text,
- over 2 million words of text parsed for predicate-argument structure,
- and 1.6 million words of transcribed spoken text annotated for speech disfluencies [2].

Its corpus is composed of texts, derived from different sources, e.g., Wall Street Journal articles, IBM computer manuals, etc., and also from transcribed telephone conversations. Data produced by the Treebank are released through the Linguistic Data Consortium (LDC)[1]. The annotation process is achieved through a two-step procedure, which involves an automatic tagging and a human correction.

PropBank (Proposition Bank) is a project of Automatic Content Extraction (ACE), which aims at creating a text corpus annotated with information on basic semantic propositions. Predicate-argument relations were added to the syntactic trees of the Penn Treebank. Thus, PropBank offers

[1] https://www.ldc.upenn.edu/.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-457-2

predicate-argument annotations, presenting a single instance, which contains information about the location of each verb, and the location and identity of its arguments [3].

III. LINGUISTIC ANALYSIS AND ONTOLOGY LEARNING

The main aim of ontology learning is retrieving from the knowledge extracted concepts and relationships among concepts. To develop adequate tools for this task, shallow and deep semantic analysis approaches are used.

Generally speaking, "the shallow semantic analysis measures only word overlap between text and hypothesis" [4]. Starting from tokenization and lemmatization of text and hypothesis, this analysis uses Web documents as a corpus and assigns to each entry inverse document frequency as a weight in the hypothesis. Thus, we have a higher score for less occurring words, which means that we assign more importance to less frequent words. Shallow analysis needs tagged corpora as training resources. This technique may be applied at both syntactic and semantic level. Shallow approach is largely used in various tasks of ontology learning:

- Term Extraction: terms are extracted using chunkers. Outputs, as nominal phrases, may be included in the basic vocabulary of the domain. Usually, in order to evaluate weight of extracted terms with respect to the corpus, statistical measures of Information Extraction (IE), such as Term Frequency for Inverse Document Frequency (TF*IDF) algorithm [5], are applied.
- Taxonomy Extraction: this task is related to the extraction of hierarchical relations among ontology classes or individuals [6] [7]. The hierarchy is usually extracted using lexical and syntactic patterns, expressed by means of regular expressions.
- Relation Extraction: using shallow parsing, it is possible to extract only limited reliable relations, i.e., simple patterns such as Noun Phrase + Verb Phrase + Noun Phrase. This analysis does not address complex sentence structures in which there are discontinued dependencies or other language ambiguities. Obviously, this limit does not allow axiom learning, obtainable only with deeper syntactic methods.

While shallow NLP covers syntactic steps as for the learning process, various methods are also applied to generate a shallow semantic parsing (semantic tagging). These methods are more useful in ontology population procedure than in learning tasks, because they govern an extraction approach relied on conceptual structures. Such approach is extremely different from the one based only on texts and syntactic NLP. Indeed, to extract entities and semantic relationships, semantic parsing requires the identification of the structures presented in the corpus. Thus, in order to discover instances of these resources, population process relies on a set of knowledge resources, such as frame, templates or roles [8]. According to [9], these resources may include role taxonomies, lists of named entities and also lexicons and dictionaries. For these reasons, shallow semantic parsing necessitates word sense disambiguation process, useful to assign to a given word the correct meaning or concept. This procedure is also applied to recognize particular semantic relationships, such as synonyms, meronyms, or antonyms using predefined patterns.

While shallow semantics may adequately respond to some ontology learning steps, the results are inadequate for more complex tasks. Shallow methods do not guarantee a fine-grained linguistic analysis. For instance, as for anaphora resolution, or quantifier scope resolution, extracting rich domain ontologies requires text processing. Considering that deep NLP allows to work not only on concepts and relations but also on axioms, such approach seems more appropriate for understanding the meaning of sentences and discourses. Indeed, if shallow methods focus only on text portions, deep ones reach a fine-grained analysis working on the whole meaning of a sentence or a discourse.

Deep methods represent a useful approach to extract representations and to infer on the basis of such representations. It means that this kind of analysis may contribute to inferencing and reasoning capabilities of machines through textual Web resources representation based on a machine-readable standard ontological language. Due to the need of applying an ontological language, in order to process textual resources, it is necessary to use grammar rules. Such set of grammar rules may be applied by a syntactic parser, "the first essential component for a deep analysis of texts" [8]. Indeed, syntactic parsing uses a set of grammar rules, known as syntactic grammars, in order to assign parse trees to sentences.

A formal language and its syntactic grammar rely also on a vocabulary, which includes all the acceptable combinations of characters of a specific alphabet. Such predefined vocabulary may be used in parsing sentences. Another way to create the lexical knowledge base useful to parse a sentence is based on training sets of hand-labeled sentences. This methodology represents the foundation of statistical parsers [10].

Parsing produces outputs represented in the form of phrase structure trees or dependency parses. "Phrase structure parses associates a syntactic parse in the form of a tree to a sentence, while dependency parses creates grammatical links between each pair of words in the sentence" [8]. By most syntactic theories, both formalizing methods are applied as complementary and not as opposite approaches. Many scholars [11] [12], also in shared tasks in Conference on Natural Language Learning (CoNLL), apply dependency parsing because it allows to model predicate-argument structures in a more intuitive way. Indeed, using predicate-argument structures for IE paradigms enables high quality IE results.

Various researches aim at establishing a correspondence between predicate-argument structure and first order predicate logic, even if this goal seems problematic. Also, according to [13], "the predicate/argument system of natural language is more complex than that of first order predicate logic".

IV. FRAMEWORK

Most of ontology learning methodologies apply syntactic parsing, based on patterns or machine learning, to improve

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-457-2

extraction of relevant structures. It means that syntactic parsing may allow a fine-grained analysis, guaranteeing also the extraction both of Atomic Linguistic Units (ALUs) and relations and axioms learning. The method applied must be adequate to the particular task we want to perform: extracting a whole ontology or only a constituents of such ontology, i.e., classes, relations or axioms.

Actually, various methodologies have been applied to increase retrieval and extraction system performance in different knowledge domains. The common aim is to process unstructured texts and, through semantic annotation procedures, formalize them in a structured representation. This step of converting texts represents the way in which we move to machine-readable language to systemize, manage and extract knowledge from the amount of data. Subtasks, involved in the formalizing process, concern entities and relations between them and their attributes. It means that in a text we have to analyze not only subjects and objects, which take part in a specific situation, that is discourse and sentence contexts, but also identify which kind of relation exists among them. Reconstructing the network of relations and attributes among entities lead us to reconstruct Aristotelian definition process of a concept. Thus, we get close to understand the meaning expressed in a text, which may be analyzed through a precise formalization of natural language, based on linguistic studies rather than on the development of stochastic algorithms. Due to such considerations, we apply Lexicon-Grammar (LG) methodologies in order to create Linguistic Resources (LRs) semantically annotated. LG, set up by Maurice Gross [14] during the '60s, is based on a language formalization achieved through a deep lexical analysis.

V. SEMANTIC TAGGING

As presented in [15], our semantic annotation process is structured into a two-step procedure: first, we tag electronic dictionary entries, and then we develop Finite State Automata/Transducers (FSA/FSTs) in order to recognize and annotate predicate-argument structures. The utility of assigning semantic labels to words is strictly linked both to the definition of semantic predicates, and to the creation of FSA/FSTs for coherent text processing. Gross' definition starts from the fact that, for each given language analyzed, LG can establish sets of lexical-syntactic structures, on the basis of the semantic features of each verb. These features are made explicit directly by the application of the rules of co-occurrence and selection restriction, through which verbs semantically select their arguments to construct acceptable simple sentences. Also, the arguments selected by each verb are given the value of attants (subjects included). Therefore, we may have semantic predicates expressing the intuitive notion of "exchange" (i.e., "Transfer Predicates" as "to give" or "to receive"), "motion" ("Movement Predicates" as "to go" or "to move") or "production" ("Creation Predicates" as "to build", "to assemble"). Each set of semantic predicates will select only and exclusively those arguments, which have with them compatible semantic roles. For instance, "Transfer Predicates" will select a "giver", an "object to transfer" and a "receiver", as in:

$\text{Max}_{(\text{giver})}$ gives a $\text{present}_{(\text{object to transfer})}$ to $\text{John}_{(\text{receiver})}$ (1)
$\text{John}_{(\text{receiver})}$ receives a $\text{present}_{(\text{object to transfer})}$ from $\text{Max}_{(\text{giver})}$ (2)

"Movement Predicates" will select an "agent of motion", eventually an "object to move", and a "locative name", as in:

$\text{Max}_{(\text{agent of motion})}$ goes to $\text{Rome}_{(\text{locative name})}$ (3)
$\text{Max}_{(\text{agent of motion})}$ moves the $\text{table}_{(\text{object to move})}$ from the living room to the $\text{kitchen}_{(\text{locative names})}$ (4)

On such basis, electronic dictionary nouns may also be labeled predicting their likelihood of becoming arguments of (a specific set of) semantic predicates. However, the list of semantic tags likely to be used is not easily identifiable, due to the polysemy of simple nouns. In fact, from the above examples, it can be seen that "agent of motion" and "creator" are sub-classes of the class "Hum", and that "object to move", and "creation" are sub-classes of the class "Conc" (concrete objects). Moreover, the words abbey and train can be selected by both "Motion" and "Creation":

Max (entered + built) the (abbey + train) (5)

and also occur as human nouns:

The (abbey + train) laughed at Max's joke (6)

On this basis, 'abbey' and 'train' could be labelled as follows:

abbey,N+FLX=APPLE+Conc
abbey,N+FLX=APPLE+Hum
abbey,N+FLX=APPLE+Loc
train,N+FLX=APPLE+Conc
train,N+FLX=APPLE+Hum
train,N+FLX=APPLE+Loc

An attempt to define a comprehensive set of semantic tags is currently in progress for the Italian DELAS-DELAF and has produced the list, presented in Table I.

TABEL I. LIST OF SEMANTIC CATEGORIES AND LABELS.

NAbb: clothing article	NLud: game/sport
NAlimE: edible substance	NMal: illness/disease
NAlimP: potable substance	NMass: mass
NAnim: animal/animate	NMat: material
NArr: (piece of) furniture	NMec: object with parts to assemble
NAst: abstract	NMeta: non-physical/metaphysical
NAtmo: weather event	NMis: unit of measure
NBot: botanical	NMon: currency
NChim: chemical	NMus: musical instrument
NColl: collective human	NNum: numeric
NConc: concrete	NPc: body-part
NCosm: cosmetic	NPcOrg: human and/or

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-457-2

	animal organism
NCreat: creation	NPsic: psychic/psychological state
NDisp: device	NQual: positive/negative quality
NEComOr: oral communication	NQuantD: defined quantifier
NEComScr: written communication	NQuantI: undefined quantifier
NEdi: construction	NSostG: gaseous-state substance
NFarm: drug or medication	NSostL: liquid-state substance
NFig: figurative	NSostS: solid-state substance
NGramm: grammatical, morphological, syntactic	NStrum: mechanical tool/object
NLin: tongue, dialect, jargon	NTmp: defined/undefined period of time/event
NLiq: non-potable substance	NUm: human
NLoc: locative	NVeic: vehicle

In terms of sets, the semantic features specified in this list overlap to a variable extent, especially with regard to all the possible subclasses of concrete nouns. For this reason, during the labeling of nouns, it will be possible to assign more than one tag to a single name.

VI. CONCLUSIONS AND FUTURE WORKS

This system of semantic classification is still in an embryonic state, but it could simplify and make even more efficient the building of NooJ FSA/FST grammars [16], allowing for verbs and nouns the insertion into nodes of one or more tags, which could be used to identify classes of words instead of single words. In this sense, it could also be possible to build large-coverage grammars for single sets of semantic predicates, as the one presented in Figure 1, which accounts for 105 simple sentences of the following kind:

(He + Max) (draws + is drawing) (a picture + pictures) (7)

(We + Paul and John) (outline + are outlining) (a drawing + drawings) (8).

REFERENCES

[1] J. Bos and K. Markert. "Combining shallow and deep NLP methods for recognizing textual entailment", in Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK, 2005, pp. 65-68.

[2] A. Taylor , M. Marcus, and B. Santorini, "The Penn TreeBank: an overview", in Treebanks, Springer Netherlands, 2003, pp. 5-22.

[3] M. Palmer, D. Gildea, and P. Kingsbury, "The Proposition Bank: A Corpus Annotated with Semantic Roles", Computational Linguistics Journal, 31:1, 2005.

[4] J. Bos and K. Markert, "Combining shallow and deep NLP methods for recognizing textual entailment", in Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK, 2005, pp. 65-68.

[5] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information processing & management 24, no. 5, 1988, pp. 513-523.

[6] S. Staab and A. Maedche. "Knowledge portals: Ontologies at work." AI magazine 22, 2001, no. 2: 63.

[7] P. Cimiano and J. Völker. "Towards large-scale, open-domain and ontology-based named entity classification." In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP). 2005, pp. 166–172.

[8] A. Zouaq and R. Nkambou. "Building domain ontologies from text for educational purposes." Learning Technologies, IEEE Transactions on no. 1, 2008, pp. 149-62.

[9] A. Giuglea and A. Moschitti, "Shallow Semantic Parsing Based on FrameNet, VerbNet and PropBank", in Proceedings of 17th European Conference on Artificial intelligence, IOS Press , 2006, pp. 563-56.

[10] D. Klein and C. D. Manning, "Accurate unlexicalized parsing", in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, 2003, pp. 423-430.

[11] T. Briscoe, and J. Carroll, "Evaluating the accuracy of an unlexicalized statistical parser on the PARC DepBank", in Proceedings of the COLING/ACL on Main conference poster sessions, Association for Computational Linguistics, 2006 pp. 41-48.

[12] S. Kübler, R. McDonald, and J. Nivre, "Dependency parsing", Synthesis Lectures on Human Language Technologies 1, no. 1, 2009, pp. 1-127.

[13] E. Luuk, "The noun/verb and predicate/argument structures", Lingua 119, no. 11, 2009, pp. 1707-1727.

[14] M. Gross, "Empiric Bases of semantic-predicate notion" (*"Les bases empiriques de la notion de prédicat sémantique*"), in Langages, 15e année, n°63, 1981, pp. 7-52.

[15] S. Vietri and M. Monteleone, "The NooJ English dictionary", in S. Koeva, S. Mesfar and M. Silberztein (eds.) Formalising Natural Language with NooJ 2013: selected papers from the NooJ 2013 International Conference, Cambridge Scholars Publishing, Newcastle upon Tyne, UK, 2013, pp. 69-86.

[16] M. Silberztein, "Formalizing languages: the NooJ approach" ("*La formalisation des langues : l'approche de NooJ*"). ISTE: London, 2015.

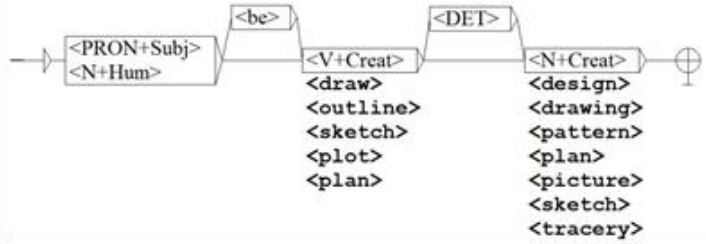

Figure 1. Sample Local Grammar for Creation of Semantic Predicates.