BOEMIE: Reasoning-based Information Extraction

Georgios Petasis

Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research (N.C.S.R.) "Demokritos"
GR-153 10, P.O. BOX 60228, Aghia Paraskevi, Athens, Greece
petasis@iit.demokritos.gr

Abstract. This paper presents a novel approach for exploiting an ontology in an ontology-based information extraction system, which substitutes part of the extraction process with reasoning, guided by a set of automatically acquired rules.

1 Introduction

Information extraction (IE) is the task of automatically extracting structured information from unstructured documents, mainly natural language texts. Due to the ambiguity of the term "structured information", information extraction covers a broad range of research, from simple data extraction from Web pages using patterns and regular grammars, to the semantic analysis of language for extracting meaning, such as the research areas of word sense disambiguation or sentiment analysis. The basic idea behind information extraction (the concentration of important information from a document into a structured format, mainly in the form of a table) is fairly old, with early approaches appearing in the 1950s, where the applicability of information extraction was proposed by the Zellig Harris for sub-languages, with the first practical systems appearing at the end of the 1970s, such as Roger Schank's systems [26, 27], which exported "scripts" from newspaper articles. The ease of evaluation of information extraction systems in comparison to other natural language processing technologies such as machine translation or summarisation, where evaluation is still an open research issue, made IE systems quite popular and led to the Message Understanding Conferences (MUC) [20] that redefined this research field.

Ontology-Based Information Extraction (OBIE) has recently emerged as a subfield of information extraction. This synergy between IE and *ontologies* aims at alleviating some of the shortcomings of traditional IE systems, such as efficient representation of domain knowledge, portability into new thematic domains, and interoperability in the era of Semantic Web [14]. Ontologies are a means for sharing and re-using knowledge, a container for capturing semantic information of a particular domain. A widely accepted definition of ontology in information technology and AI community is that of "a formal, explicit specification of a shared

conceptualization" [28, 10], where "formal implies that the ontology should be machine-readable and shared that it is accepted by a group or community" [4]. According to [30], an ontology-based information extraction system is a system that "processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies". This definition suggests that the main differences between traditional IE systems and OBIEs are: a) OBIEs present their output using ontologies, and b) OBIEs use an information extraction process that is "guided" by an ontology. In all OBIE systems the extraction process is *guided* or *driven* by the ontology to extract things such as classes, properties and instances [30], in a process known as *ontology population* [22].

However, the way the extraction process is guided by an ontology in all OBIEs has not changed much with respect to traditional information extraction systems. According to a fairly recent survey [30], OBIEs do not employ new extraction methods, but they rather employ existing methods to identify the components of an ontology. Current research on the field researches the development of "reusable extraction components" that are tied to ontology portions that are able to identify and populate [29, 11]. In this paper we propose an alternative approach that tries to *minimise* the use of traditional information extraction components, and substitute their effect with *reasoning*. The motivation behind the work presented in this paper is to propose a new "kind" of ontology-based information extraction system, which integrates further ontologies and traditional information extraction approaches, through the use of *reasoning* for "guiding" the extraction process, instead of heuristics, rules, or machine learning. The proposed approach splits a traditional OBIE in two parts, the first part of which deals with the gathering of evidence from documents (in the form of ontology property instances and relation instances among them), while the second part employs reasoning to interpret the extracted evidence, driven by plausible explanations for the observed relations. Thus, the innovative aspects of the presented approach include a) the use of ontology through reasoning as a substitute for the embedded knowledge usually found in the extraction components of OBIEs, b) a proposal of how reasoning can be applied for extracting information from documents, and c) an approach for inferring the required interpretation rules even when the ontology evolves with the addition of new concepts and relations.

The rest of this paper is organised as follows: In section 2 related work is presented in order to place our approach within the current state-of-the-art. In section 3 the proposed approach is presented, detailing both the interpretation process and the automatic reasoning rule acquisition. Finally, section 4 concludes this paper and outlines interesting directions for further research.

2 Related Work

Ontology-based information extraction has recently emerged as a subfield of information extraction that tries to bring together traditional information extraction and ontologies, which provide formal and explicit specifications of con-

ceptualizations, and acquire a crucial role in the information extraction process. A set of recent surveys have been presented that analyse the state-of-art in the research fields of OBIEs [13, 14, 30] and ontology learning/evolution [23, 22], a relevant research field since many OBIE systems also perform ontology evolution/learning. OBIE systems can be classified according to the way they acquire the ontology to be used for information extraction. One approach is to consider the ontology as an input to the system: The OBIE is guided by a manually constructed ontology or ~~form~~ an "off-the-shelf" ontology. Most OBIE systems appear to adopt this approach [30]. Such systems include SOBA [5, 3], KIM [24, 25] the implementation by Li and Bontcheva [17] and PANKOW [7], Artequact [15, 2, 1]. The other approach is to construct an ontology as a part of the information extraction process, either starting from scratch or by evolving an initial, seed ontology. Such systems include Text-To-Onto [18], the implementation by Hwang [12], Kylin [31], the work by Maedche et al. [19], the work of Dung and Kameyama [8]. However, all the aforementioned systems employ traditional information extraction methods to identify elements of the ontology, and none attempts to employ reasoning, as the work presented in this paper suggests.

3 The BOEMIE approach

The work presented in this paper has been developed in the context of the BOEMIE project. It advocates an ontology-driven multimedia content analysis, i.e. semantics extraction from images, video, text, audio/speech, through a novel synergistic method that combines multimedia extraction and ontology evolution in a bootstrapping fashion. This method involves, on one hand, the continuous extraction of knowledge from multimedia content sources in order to populate and enrich the ontologies and, on the other hand, the deployment of these ontologies to enhance the robustness of the multimedia information extraction system. More details about BOEMIE can be found in [6, 22].

As already mentioned, the proposed approach splits a traditional OBIE in two parts, the first part of which deals with the gathering of evidence from documents (in the form of ontology property instances and relation instances among them), while the second part employs reasoning to interpret the extracted evidence, driven by plausible explanations for the observed relations. As a result, the typical extraction process in also split in two phases: "low-level analysis" (where traditional extraction techniques such as machine learning are used) and "semantic interpretation", where analysis' results are explained, according to the ontology, through reasoning. Each of the two phases identifies different elements of the ontology, whose elements are also split in two groups, the "mid-level concepts" (MLCs - identified by low-level analysis), and the "high-level concepts" (HLCs), which are identified through semantic interpretation.

The implications of this separation are significant: the low-level analysis cannot assume that a Person/Athlete/Journalist has been found in a multimedia document, just because a name has been identified. Instead the low-level analysis reports that a name, an age, a nationality, a performance, etc. has been found,

and reports how all these are related through binary relations, extracted from modality-specific information (i.e. linguistic events for texts, spatial relations for images/videos, etc.). The identification of Person/Athlete/Journalist instances is done through reasoning, using the ontology and the reasoning (interpretation) rules, as low-level analysis cannot know how the Person or Athlete concepts are defined in the ontology (i.e. what their properties/axioms/restrictions are). In essence, BOEMIE proposes a novel approach for constructing an OBIE, by keeping the named-entity extraction phase from traditional IE systems, modifying relation extraction to reflect modality-specific relations at the ontological level, and implementing the remaining phases of traditional IE systems through reasoning. For example, low-level analysis of an image is responsible for reporting only that a few tenths of faces have been detected (i.e. the faces of athletes and the audience – represented as MLC instances), along with a human body (i.e. the body of an athlete – represented as an MLC instance), a pole, a mattress, two vertical bars, a horizontal bar, etc. (all these are instances of MLCs). After MLC instances have been identified, the low-level analysis is expected to identify relational information about these MLC instances. For example, the low-level analysis is expected to identify that a specific face is adjacent to a human body and both are adjacent to the pole and the horizontal bar. The low-level analysis is expected to report the extracted relational information through suitable binary relations between each pair of related MLC instances. On the other hand, the low-level is not expected to interpret its findings and hypothesise instances of HLC instances, such as the existence of athletes and their number. It is up to the second phase, the semantic interpretation, to identify how many athletes are involved (each one represented as instance of the "Athlete" HLC), and to interpret the scene shown in the image as an instance of the "Pole Vault" HLC concept, effectively explaining the image.

3.1 Definitions

The approach presented in this paper organises the ontology into four main ontological modules, the "low-level features", the "mid-level concepts", the "high-level concepts", and the "interpretation rules", which are employed through reasoning in order to provide one or more "interpretations" of a multimedia document.

Definition 1 (low-level features). *Low-level features are concepts related to the decomposition of a multimedia document (i.e. the description of an HTML page into text, images or other objects), and concepts that describe surface forms on single modality documents, such as segments in text and audio documents, polygons in image and video frames, etc.*

Definition 2 (Mid-Level Concept (MLC)). *Mid-level concepts are concepts that can be materialised (i.e. have surface forms) on documents of a single modality. Anything that can be extracted by an OBIE that has a surface form on a document, is an MLC.*

For example, the names of persons, locations, etc. in texts, the faces, bodies of persons in images and the sound events (i.e. applauses) in audio tracks are all MLC concepts. The BOEMIE OBIE extracts only instances of MLCs *(MLCis)* and relations (i.e. spatial) among them.

Definition 3 (High-Level Concept (HLC)). *High-level concepts are compound concepts formed from MLCs. HLCs cannot be directly identified in a multimedia document, as they cannot be associated with a single surface form (i.e. segment).*

For example, the concept "Person" is an HLC, that groups several MLCs (properties), such as "PersonName", "Age", "Nationality", "PersonFace", "PersonBody", etc. Instances of HLCs *(HLCis)* in the BOEMIE OBIE are identified through reasoning over MLC instances (MLCis) in the ontology, guided by a set of rules, in a process known as "interpretation".

Definition 4 (interpretation). *Interpretation is the identification of one or more HLC instances (HLCis) in a multimedia document.*

An OBIE can have identified several MLC instances (MLCis) and relations between them in a multimedia document. If these MLC instances satisfy the axioms of the ontology and the interpretation rules are able to generate one or more HLC instances (HLCis), then this multimedia document is considered as *interpeted* (or *explained*) by the ontology, with the HLC instances (HLCis) constituting the *interpretation* of the document. If the same MLC instances (MLCis) are involved in more than one HLC instances (HLCis) of the same HLC, then the document is considered to have *multiple interpretations*, usually due to *ambiguity*.

3.2 Semantic Extraction

The extraction engine is responsible for extracting instances of concept descriptions that can be directly identified in corpora of different modalities. These concept descriptions are mid-level concepts (MLCs). For example, in the textual modality the name or the age of a person is an MLC, as instances of these concepts are associated directly with relevant text segments. On the other hand, the concept person is not an MLC, as it is a "compound", or "aggregate" concept in such a way that instances of this concept are related to instances of name, age, gender or maybe instances of other compound concepts. Compound concepts are referred to as high-level concepts (HLCs), and instances of such concepts cannot be directly identified in a multimedia document, and thus associated with a content segment. Thus, such instances and also relationships between these instances have to be hypothesized. In particular, this engine implements a modular approach [13] that comprises the following three level of abstraction: 1. The low-level analysis, which includes a set of modality-specific (image, text, video, audio) content analysis tools. 2. A modality-specific semantic interpretation engine. 3. A fusion engine, which combines interpretations from each modality[1].

[1] The fusion engine will not be described in this paper, as it is similar to the semantic interpretation engine. More information can be found at [6, 22].

The first two levels implement ontology-driven, modality-specific information extraction, while the last one fuses the information obtained from the previous levels of analysis. The first level involves the identification of "primitive" concepts (MLCs), as well as instances of binary relations amongst them. The second level involves the semantic interpretation engine, responsible for hypothesizing instances of high-level concepts (HLCs) representing the interpretation of (parts of) a document. Semantic interpretation operates on the instances of MLCs and relations between them extracted by the information extraction engine. The goal of semantic interpretation is to explain why certain instances of MLCs are observed in certain relations according to the background knowledge (domain ontology and a set of interpretation rules) [9], by creating instances of high-level concepts and relating these instances. Semantic interpretation is performed through calls to a non-standard reasoning service, known as explanation derivation via abduction. The semantic interpretation is performed on the extracted information (MLC/relation instances) from a single modality in order to form modality-specific HLC instances. The fact that content analysis is separated from semantic interpretation, along with the fact that semantic interpretation is performed through reasoning using rules from the ontology, allows single-modality extraction to be adaptable to changes in the ontology.

Once a multimedia document has been decomposed into single-modality elements and each element has been analysed and semantically interpreted separately, the various interpretations must be fused into one or more alternative interpretations of the multimedia document as a whole. This process is performed at a third level, where the modality-specific HLC instances are fused in order to produce HLC instances that are not modality-specific, and contain information extracted from all involved modalities. Fusion is also formalized as explanation generation via abductive reasoning.

Example: the OBIE for the text modality The low-level analysis system implemented in the context of BOEMIE exploits the infrastructure offered by the Ellogon[2] platform [21], and the Conditional Random Fields [16] machine learning algorithm, in order to build an adaptable named-entity recognition and classification (NERC) system, able to identify ~~instances of MLCs~~ (MLCis) and relations between MLCis. Both NERC and relation extraction components operate in a supervised manner, using MLC instances that populate the (seed or evolved) ontology as training material (whose surface forms are available through their low-level features). The fact that both components use the populated ontology as training source, allows them to adapt to ontology changes, and improve their extraction performance over time, as the ontology evolves. The performance of the NERC and relation extraction components has been measured to about 85% and 70% (F-measure), in the thematic domain of athletics, involving news items and biographies from official sites like IAAF[3] (International Association

[2] http://www.ellogon.org
[3] http://www.iaaf.org/

of Athletics Federations). More details about the low-level analysis system for the text modality can be found in [13].

The modality specific interpretation engine (not only for text, but for all modalities) is a process for generating instances of HLCs, by combining instances of MLCs, through reasoning over instances. Abduction is used for this task, a type of reasoning where the goal is to derive explanations (causes) for observations (effects). In the framework of this work we regard as explanations the high-level semantics of a document, given the middle-level semantics, that is, we use the extracted MLCis in order to find HLCis [9]. The reasoning process is guided by a set of rules, which belong into two kinds, deductive and abductive. Assuming a knowledge base, $\Sigma = (T, A)$ (i.e. an ontology), and a set of assertions Γ, (i.e. the assertions of the semantic interpretation of a document), abduction tries to derive all sets of assertions (interpretations) Δ such as $\Sigma \cup \Delta \models \Gamma$, while the following conditions must be satisfied: (a) $\Sigma \cup \Delta$ is satisfiable, and (b) Δ is a minimal explanation for Γ, i.e. there exists no other explanation Δ' (not equivalent to Δ) that $\Sigma \cup \Delta' \models \Delta$ holds. For example, assuming the following ontology Σ (containing both a "terminological component" – TBox, and a set of rules):

$$
\begin{aligned}
Jumper &\sqsubseteq Human \\
Pole &\sqsubseteq SportsEquipment \\
Bar &\sqsubseteq SportsEquipment \\
Pole \sqcap Bar &\sqsubseteq \bot \\
Pole \sqcap Jumper &\sqsubseteq \bot \\
Jumper \sqcap Bar &\sqsubseteq \bot \\
JumpingEvent &\sqsubseteq \exists_{\leq 1} hasParticipant.Jumper \\
PoleVault &\sqsubseteq JumpingEvent \sqcap \exists hasPart.Pole \sqcap \exists hasPart.Bar \\
HighJump &\sqsubseteq JumpingEvent \sqcap \exists hasPart.Bar \\
near\,(Y,Z) &\leftarrow PoleVault\,(X)\,, hasPart\,(X,Y)\,, Bar\,(Y)\,, \\
&\quad\; hasPart\,(X,W)\,, Pole\,(W)\,, hasParticipant\,(X,Z)\,, Jumper\,(Z) \\
near\,(Y,Z) &\leftarrow HighJump\,(X)\,, hasPart\,(X,Y)\,, Bar\,(Y)\,, \\
&\quad\; hasParticipant\,(X,Z)\,, Jumper\,(Z)
\end{aligned}
$$

And a document (i.e. an image) describing a pole vault event, whose analysis results Γ contain instances of the MLCs "Pole", "Human", "Bar" and a relation that the human is near the bar:

$$
\begin{aligned}
pole_1 &: Pole \\
human_1 &: Human \\
bar_1 &: Bar \\
(bar_1, human_1) &: near
\end{aligned}
$$

The interpretation process splits the set of analysis assertions Γ into two subsets: (a) Γ_1 (bona fide assertions): $\{pole_1 : Pole, human_1 : Human, bar_1 : Bar\}$,

which are assumed to be true by default, and (b) Γ_2 (fiat assertions): $\{(bar_1, human_1 : near)\}$, containing the assertions aimed to be explained. Since Γ_1 is always true, $\Sigma \cup \Delta \models \Gamma$ can be expressed as $\Sigma \cup \Gamma_1 \cup \Delta \models \Gamma_2$. Then, a query Q_1 is formed from each fiat assertion (Γ_2), such as $Q_1 := \{() | near\,(bar_1, human_1)\}$. Executing the query, a set of possible explanations (interpretations) is retrieved:

$$\begin{aligned}
\Delta_1 = \{&NewInd_1 : PoleVault, (NewInd_1, bar_1) : hasPart,\\
&(NewInd_1, NewInd_2) : hasPart, NewInd_2 : Pole,\\
&(NewInd_1, human_1) : hasParticipant, human_1 : Jumper\}\\
\Delta_2 = \{&NewInd_1 : PoleVault, (NewInd_1, bar_1) : hasPart,\\
&(NewInd_1, pole_1) : hasPart, (NewInd_1, human_1) : hasParticipant,\\
&human_1 : Jumper\}\\
\Delta_3 = \{&NewInd_1 : HighJump, (NewInd_1, bar_1) : hasPart,\\
&(NewInd_1, human_1) : hasParticipant, human_1 : Jumper\}
\end{aligned}$$

Each interpretation is scored, according to a heuristic based on the number of hypothesized entities and the number of involved Γ_1 assertions used, and the best scoring interpretations are kept. For the example interpretation shown above, Δ_2 is the best scoring explanation, as Δ_1 has an excessive hypothesized entity ($NewInd_2$), and Δ_3 does not use the "Pole" instance from Γ_1. More details about interpretation through abduction can be found in [6] and [9].

3.3 The Role of Interpretation Rules

Rules are considered part of the ontology TBox and their role is to provide guidance to the interpretation process. Their main responsibility is to provide additional knowledge on how analysis results (specified through MLCis and relations between MLCis) can be mapped into HLCis within a single modality, and how HLCis from various modalities can be fused. As a result, rules can be split in two categories: rules for semantic interpretation, and rules for fusion. Both categories follow the same design pattern for rules: each rule is built around a specific instance or a relation between two instances in the left hand side (LHS) of the rule, followed by a set of statements or restrictions in the right hand side (RHS) of the rule. When a rule is applied by the semantic interpretation engine, instances can be created to satisfy the rule, either for concepts/relations of the LHS (forward rules) or for concepts in the RHS (backward rules).

Forward rules Forward rules perform an action (usually the addition of a relation between two instances) described in the LHS of the rule, if the restrictions contained in the RHS have been satisfied. For example, consider the following ABox fragment:

$$\begin{aligned}
(personName_1, \text{“}JaroslavRybakov\text{”}) &: hasValue\\
(ranking_1, \text{“1”}) &: hasValue
\end{aligned}$$

$$(person_1, personName_1) : hasPersonName$$
$$personName_1 : PersonName$$
$$person_1 : Person$$
$$ranking_1 : Ranking$$
$$(personName_1, ranking_1) : personNameToRanking$$

This ABox fragment describes the situation where the semantics extraction engine has identified two MLCis, a person name ("Jaroslav Rybakov") and a ranking ("1"), connected with the "personNameToRanking" relation. Also, a "Person" instance exists that relates only to the "PersonName" instance, but not to the "Ranking" instance. Despite the fact that the $personName_1$ MLCi is related to the $ranking_1$ MLCi, the $person_1$ HLCi that aggregates $personName_1$ is not related to the "Ranking" instance. In order for the "Person" instance to be related to the "Ranking" instance, a forward rule like the following one must be present during interpretation:

$$personToRanking(X, Z) \leftarrow Person(X), PersonName(Y), hasPersonName(X, Y), personNameToRanking(Y, Z)$$

This rule can be interpreted as follows: if a "Person" instance X and a "PersonName" instance Y are found connected with a $hasPersonName(X, Y)$ relation, and a relation "personNameToRanking" exists between the "PersonName" Y and any instance Z, then add a relation between the "Person" instance X and the instance Z. The fact that the rule is applied in a forward way, suggests that all restrictions in the RHS have to be met, for the relation "personToRanking" on the LHS to be added in an ABox.

Backward rules Backward rules on the other hand assume that the restriction described by the LHS is already satisfied by the ABox (i.e. instances and relations exist in the ABox), and that the action involves the addition of (one or more) missing instances or relations to satisfy the RHS. Consider for example the following ABox fragment from the image modality:

$$personBody_1 : PersonBody$$
$$personFace_1 : PersonFace$$
$$(personBody_1, personFace_1) : isAdjacent$$

This ABox fragment describes two MLCis (a person face and a person body) that are found adjacent inside an image. Also, suppose that the TBox contains a backward rule like the following one:

$$isAdjacent(Y, Z) \leftarrow Person(X), PersonBody(Y), PersonFace(Z), hasPart(X, Y), hasPart(X, Z)$$

This rule roughly suggests that if a person face and a person body instances are aggregated by a person instance (and thus both body parts are related to the person instance with the "hasPart" relation), then the two body parts must be adjacent to each other. However, since the relation $isAdjacent(personBody_1, personFace_1)$ already exists in the ABox and the rule is a backward one, it will try to hypothesise a "Person" instance X, and aggregate the two body parts.

3.4 Rules for Semantic Interpretation

One domain of rules application is the semantic interpretation of the results obtained from the low level analysis, performed on multimedia resources. During this interpretation process, the MLCis and the relations among MLCis are examined, in order to aggregate the MLCis into HLCis. Then, relations that hold between MLCis are promoted to the HLCis that aggregate the corresponding MLCis. Finally, an iterative process starts, which tries to aggregate the HLCis into other HLCis and again promote the relations, until no other instances can be added to an ABox. As a result, two types of rules are required during interpretation: rules that aggregate concept instances (either MLCis or HLCis) into instances of HLCs, and rules that promote relations. However, not all relations must be promoted: only relations that hold between a property instance of an HLCi and an instance that is not a property of the HLCi should be promoted to the HLCi. The aggregation of instances into HLCis is performed with the help of *backward rules*[4], while the promotion of relations from properties to the aggregating HLCi is performed with *forward rules*.

3.5 Acquiring Rules

When the ontology changes (i.e. through the addition of a new concept) the interpretation rules must be modified accordingly. We tried to automate this task by monitoring ontological changes: the actions performed by an ontology expert to the ontology are monitored and reflected to the interpretation rules, following a transformation based approach. Considering as input what an ABox can contain without the current concept definition available, and as output the instances that can be generated from the concept if defined, the rule generation approach tries to find a set of rules that can transform the input into the desired output. In order to perform this transformation, the transformation is split into a set of more primitive "operations" that can be easily transformed into rules.

Assuming the set of all possible concepts C, the set of all possible relations R, a set of predefined operations O on a single concept $c \in C$, and a modification M over c, where $M = \{m_i\,(c, c_i, r_i)\}_1^N$, $m_i \in O$, $c_i \in C$, $c_i \neq c$, $r_i \in R$, the target is to calculate a rule set $S = \{r_i\}_1^N$, $r_i = T_{m_i}\,(c, c_i, r_i)$, that corresponds to the modification M. T_{m_i} is a function that transforms a hypothesized initial state (c', c_i, r_i') to the desired state (c, c_i, r_i) for modification m_i,

[4] Backward rules imply the use of abduction to hypothetize instances not contained in the original ABox.

$T_{m_i} : (c', c_i, r'_i) \rightarrow (c, c_i, r_i)$, $c' \in C$, $r' \in R$. Each function T_{m_i} depends not only on m_i and the two states, but also on the interpretation engine and reasoner in use. Since the objective of rules generation was to eliminate manual supervision, a pattern based approach was selected for representing each T_{m_i}. Each pattern is responsible for generating the required rules from transforming the initial state (c', c_i, r'_i) to a final state (c, c_i, r_i) for each operation in O, possibly biased towards the specific interpretation model and thus reasoner in use.

Operations over a Single Concept A set of predefined operations O has been defined that captures all modifications that can be performed on a concept c within the BOEMIE system. This set contains the following operations:

- Definition of a new MLC c: This operation reflects the addition of a new MLC to the ontology TBox, an action that is not associated with the modification of the rules associated with the TBox. For this operation, $T = \{\}$.
- Definition of a new HLC c that aggregates a single concept c': This operation describes the action of the definition of an HLC based on the presence of either an MLC or an HLC. Typical usage of this operation is when a new HLC c has been defined that aggregates another concept c', and c' is enough to define this concept c. In such a case, it is assumed that during interpretation an instance of c should be created for every instance of c' found in an ABox. Thus the set of rules T should create an instance of c for every instance of c'. Example of this operation is the definition of "Person" (c) that aggregates either "PersonName" or "PersonFace" (c').
- Addition of a single concept c' to an existing HLC c: This operation deals with the extension of an existing HLC c with a concept c', i.e. when adding a new property to an existing HLC. In such a case, T should contain rules that aggregate instances of concept c' with instances of concept c, and promote all relations between the instance of c' and instances not aggregated by the instance of c to the c instance. Examples include the extension of "Person" with properties like "Age", "Gender", or "PersonBody" and the "SportsEvent" with "Date", or "Location".
- Removal of a single concept c' from an HLC c: This operation handles property removals from HLCs. The rule set T is identical to the operation of adding a property to an HLC, with the difference that each rule in T is located and removed from the TBox rules, instead of extending it.
- Removal of HLC c that aggregates a single concept c': Again, this operation is the negation of creating a new HLC that aggregates a single concept operation. Thus, the rule set T is identical between the two operations, but this operation causes the removal of all rules in T from the TBox.
- Removal of an MLC c: Similar to the addition of a new MLC operation, this operation has no effect on the TBox rule set, i.e. no rules are removed.

Rule templates for concept definition operations In this subsection the templates for generating rules are described, for the operators that do not have

an empty set T, and are not related to removals, which share the same T with the corresponding addition operations.

Definition of a new HLC c that aggregates a single concept c' The rule set T during the definition of a new HLC c from a concept c' should contain rules that create instances of c from instances of c' found in the ABox of a multimedia resource. In the interpretation model used in BOEMIE, this can be accomplished by a single backward rule, which can be described with the following pattern:

$$\langle c' \rangle (X) \leftarrow \langle c \rangle (Y), has \langle c' \rangle (Y, X)$$

For example, if c is "Person" and c' is "PersonName", the following rule can be generated from this pattern:

$$PersonName(X) \leftarrow Person(Y), hasPersonName(Y, X)$$

Addition of a single concept c' to an existing HLC c The rule set T during the addition of a property c' to an HLC c should contain rules that relate instances of c with instances of c' found in the ABox of a multimedia resource. In addition, it should contain rules that promote the relations of a c' instance with all instances not aggregated by c onto the c instance. This operation reflects an action performed on the definition of concept c, from which the "final" state (c, c', r) is known. The state (c, c', r) is the part of the concept definition that relates to how c aggregates c'. For example, if "Person" in the image modality is defined as having only a single property $(hasPersonFace : PersonFace)$, and the operation is to extend it also with "PersonBody" through the role "hasPersonBody", then $(c, c', r) = (Person, PersonBody, hasPersonBody)$. According to the adopted interpretation model, c' can be aggregated with c only if c' is related with any property of c. If c'', $c'' \neq c'$ is an aggregated by c concept, then an "initial" state (c'', c', r'') is hypothesized, relating c' with c'' through the relation r''. Continuing the example, since "Person" has a single aggregated concept, only one initial state can be hypothesized, i.e. $(c'', c', r'') = (PersonFace, PersonBody, isAdjacent)$. Once both initial and final states have been decided, then a rule pattern can be defined to transform the initial into the final state. In the interpretation model used within BOEMIE, this can be accomplished by a single backward rule, which can be described with the following pattern:

$$\begin{aligned} \langle r'' \rangle (Y, Z) \leftarrow & \langle c \rangle (X), has \langle c'' \rangle (X, Y), \langle c'' \rangle (Y), \\ & \langle r \rangle (X, Z), \langle c' \rangle (Z) \end{aligned}$$

Applied to our example, this pattern will lead to the following rule:

$$\begin{aligned} isAdjacent(Y, Z) \leftarrow & Person(X), hasPersonFace(X, Y), PersonFace(Y), \\ & hasPersonBody(X, Z), PersonBody(Z) \end{aligned}$$

This rule can relate instances of "PersonBody" to instances of "Person", already related to instances of "PersonFace". The same process should be repeated for all possible initial states that can be found for concept c.

However these are not the only rules that should be added in set T. Each relation w defined in the TBox that can have as subject concepts c and c', must be promoted from c' to c. This can be accomplished with forward rules that can be generated by the following pattern:

$$\langle w \rangle (X, Z) \leftarrow \langle c \rangle (X), \langle r \rangle (X, Y), \langle c' \rangle (Y), \langle w \rangle (Y, Z)$$

Please note that in this pattern no type is specified for variable Z, allowing Z to take as value instances of any concept that is in the range of the relation $\langle w \rangle$. Assuming $w = isNear$, this pattern can lead to the following rule:

$$isNear(X, Z) \leftarrow Person(X), hasPersonBody(X, Y), PersonBody(Y), isNear(Y, Z)$$

The rule set T must be extended with a single rule of the above form for each w that can be found in the ontology TBox.

4 Conclusions

In this paper we have presented a novel approach for exploiting an ontology in an ontology-based information extraction system, which substitutes part of the extraction process with reasoning, guided by a set of automatically acquired rules. Innovative aspects of the presented framework include the use of reasoning in the construction of an ontology-based information extraction system that can adapt to changes in the ontology and the clear distinction between concepts of the low-level analysis (MLCs), and the semantic interpretation (HLCs). An interesting future direction is the investigation of how reasoning can be better applied on modalities involving the dimension of time, such as video. In BOEMIE a simple approach has been followed regarding the handling of time sequences, where extracted real objects or events were grounded to timestamps, and artificial relations like "before" and "after" were added. Nevertheless, an enhancement that maintains the temporal semantics from the perspective of reasoning will be an interesting addition.

Acknowledgments.

This work has been partially funded by the BOEMIE Project, FP6-027538, 6^{th} EU Framework Programme.

References

1. Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., Shadbolt, N.R.: Automatic ontology-based knowledge extraction from web documents. IEEE Intelligent Systems 18(1), 14–21 (Jan 2003), http://dx.doi.org/10.1109/MIS.2003.1179189

2. Alani, H., Kim, S., Millard, D.E., Weal, M.J., Lewis, P.H., Hall, W., Shadbolt, N.: Automatic extraction of knowledge from web documents. In: Workshop on Human Language Technology for the Semantic Web and Web Services, 2 nd Int. Semantic Web Conf. Sanibel Island (2003)
3. Buitelaar, P., Cimiano, P., Frank, A., Hartung, M., Racioppa, S.: Ontology-based information extraction and integration from heterogeneous data sources. Int. J. Hum.-Comput. Stud. 66(11), 759–788 (Nov 2008), http://dx.doi.org/10.1016/j.ijhcs.2008.07.007
4. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications, Frontiers in Artificial Intelligence and Applications Series, vol. 123. IOS Press, Amsterdam (7 2005)
5. Buitelaar, P., Siegel, M.: Ontology-based information extraction with soba. In: In: Proc. of the International Conference on Language Resources and Evaluation (LREC). pp. 2321–2324 (2006)
6. Castano, S., Peraldi, I.S.E., Ferrara, A., Karkaletsis, V., Kaya, A., Möller, R., Montanelli, S., Petasis, G., Wessel, M.: Multimedia Interpretation for Dynamic Ontology Evolution. Journal of Logic and Computation 19(5), 859–897 (2009)
7. Cimiano, P., Handschuh, S., Staab, S.: Towards the self-annotating web. In: Proceedings of the 13th international conference on World Wide Web. pp. 462–471. WWW '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/988672.988735
8. Dung, T.Q., Kameyama, W.: Ontology-based information extraction and information retrieval in health care domain. In: Proceedings of the 9th international conference on Data Warehousing and Knowledge Discovery. pp. 323–333. DaWaK'07, Springer-Verlag, Berlin, Heidelberg (2007), http://dl.acm.org/citation.cfm?id=2391952.2391991
9. Espinosa, S., Kaya, A., Melzer, S., Möller, R.: On ontology based abduction for text interpretation. In: Gelbukh, A. (ed.) Proc. of 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008). pp. 194–205. No. 4919 in LNCS, Springer (2008)
10. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum.-Comput. Stud. 43(5-6), 907–928 (Dec 1995), http://dx.doi.org/10.1006/ijhc.1995.1081
11. Gutierrez, F., Wimalasuriya, D.C., Dou, D.: Using information extractors with the neural electromagnetic ontologies. In: Meersman, R., Dillon, T.S., Herrero, P. (eds.) OTM Workshops. Lecture Notes in Computer Science, vol. 7046, pp. 31–32. Springer (2011)
12. Hwang, C.H.: Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. In: Franconi, E., Kifer, M. (eds.) KRDB. CEUR Workshop Proceedings, vol. 21, pp. 14–20. CEUR-WS.org (1999)
13. Iosif, E., Petasis, G., Karkaletsis, V.: Ontology-Based Information Extraction under a Bootstrapping Approach. In: Pazienza, M.T., Stellato, A. (eds.) Semi-Automatic Ontology Development: Processes and Resources, chap. 1, pp. 1–21. IGI Global, Hershey, PA, USA (April 2012)
14. Karkaletsis, V., Fragkou, P., Petasis, G., Iosif, E.: Ontology based information extraction from text. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds.) Knowledge-Driven Multimedia Information Extraction and Ontology Evolution. Lecture Notes in Computer Science, vol. 6050, pp. 89–109. Springer (2011)
15. Kim, S., Alani, H., Hall, W., Lewis, P., Millard, D., Shadbolt, N., Weal, M.: Artequakt: Generating tailored biographies from automatically annotated fragments

from the web. In: Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM?02), the 15th European Conference on Artificial Intelligence, (ECAI?02). vol. -, pp. 1–6 (2002), http://eprints.soton.ac.uk/256913/, event Dates: July 21-26
16. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
17. Li, Y., Bontcheva, K.: Hierarchical, perceptron-like learning for ontology-based information extraction. In: Proceedings of the 16th international conference on World Wide Web. pp. 777–786. WWW '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1242572.1242677
18. Maedche, A., Maedche, E., Staab, S.: The text-to-onto ontology learning environment. In: Software Demonstration at ICCS-2000 - Eight International Conference on Conceptual Structures (2000)
19. Maedche, A., Neumann, G., Staab, S.: Intelligent exploration of the web. chap. Bootstrapping an ontology-based information extraction system, pp. 345–359. Physica-Verlag GmbH, Heidelberg, Germany, Germany (2003), http://dl.acm.org/citation.cfm?id=941713.941736
20. Marsh, E., Perzanowski, D.: Muc-7 evaluation of ie technology: Overview of results. In: Proceedings of the Seventh Message Understanding Conference (MUC-7). http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html (1998)
21. Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., Spyropoulos, C.D.: Ellogon: A New Text Engineering Platform. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002). pp. 72–78. European Language Resources Association, Las Palmas, Canary Islands, Spain (May 29–31 2002)
22. Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E.: Ontology Population and Enrichment: State of the Art. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds.) Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Lecture Notes in Computer Science, vol. 6050, pp. 134–166. Springer Berlin / Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-20795-2_6
23. Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E.: Ontology Population and Enrichment: State of the Art. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds.) Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Lecture Notes in Computer Science, vol. 6050, pp. 134–166. Springer Berlin / Heidelberg (2011)
24. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M.: Kim - semantic annotation platform. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) International Semantic Web Conference. Lecture Notes in Computer Science, vol. 2870, pp. 834–849. Springer (2003)
25. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A.: Kim - a semantic platform for information extraction and retrieval. Natural Language Engineering 10(3-4), 375–392 (2004)
26. Schank, R.C., Abelson, R.P.: Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures. L. Erlbaum, Hillsdale, NJ (1977)
27. Schank, R.C., Kolodner, J.L., DeJong, G.: Conceptual information retrieval. In: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (SIGIR '80). pp. 94–116. Cambridge, UK (1980)

28. Studer, R., Benjamins, R., Fensel, D.: Knowledge engineering: Principles and methods. Data & Knowledge Engineering 25(1-2), 161–198 (März 1998)
29. Wimalasuriya, D.C., Dou, D.: Components for information extraction: ontology-based information extractors and generic platforms. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 9–18. CIKM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1871437.1871444
30. Wimalasuriya, D.C., Dou, D.: Ontology-based information extraction: An introduction and a survey of current approaches. J. Inf. Sci. 36(3), 306–323 (Jun 2010), http://dx.doi.org/10.1177/0165551509360123
31. Wu, F., Weld, D.S.: Automatically refining the wikipedia infobox ontology. In: Proceedings of the 17th international conference on World Wide Web. pp. 635–644. WWW '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1367497.1367583