On Ontology Based Abduction For Text Interpretation ⋆

Irma Sofia Espinosa Peraldi, Atila Kaya, Sylvia Melzer, Ralf Möller

Hamburg University of Technology,
Institute for Software Systems,
Hamburg, Germany
sofia.espinosa@tuhh.de, at.kaya@tuhh.de,
sylvia.melzer@tuhh.de, r.f.moeller@tuhh.de

Abstract. Text interpretation can be considered as the process of extracting deep-level semantics from unstructured text documents. Deep-level semantics represent abstract index structures that enhance the precision and recall of information retrieval tasks. In this work we discuss the use of ontologies as valuable assets to support the extraction of deep-level semantics in the context of a generic architecture for text interpretation.

1 Introduction

The growing amount of unstructured electronic documents is a problem found in proprietary as well as in public repositories. In this context, the web is a representative example where the need of logic-based information retrieval (IR) to enhance precision and recall is evident. Logic-based IR means the retrieval of unstructured documents with the use of abstract terms that are not directly readable from the surface of the text, but only between its lines. For example, *Chocolate Cake Recipe* is an abstract term for the following text:

> *Yield: 10 Servings, 5 oz. semisweet chocolate (chopped), 3 oz. unsweetened chocolate (chopped), 1/4 lb. (8 Tbs.) unsalted butter, 1/4 cup all-purpose flour, 4 eggs at room temperature, ...*

Relational index structures are crucial for IR. Therefore, the task of defining the necessary index structures for abstract terms to allow logic-based IR is unavoidable. In our work, the necessary structures for logic-based IR are called *deep-level semantics* and the process of extracting deep-level semantics from unstructured text documents is understood as *text interpretation*. In the course of the work presented here, we will highlight that a feasible architecture (see Figure 1) to enable the automatic extraction of deep-level semantics from large-scale corpora can be achieved through:

⋆ This paper has been partially supported by the BOEMIE Project, contract FP6-027538, under the 6th EU Framework Programme. The example ontologies used in this paper are based on the ontologies developed in BOEMIE, and practical input to the interpretation algorithms investigated in this paper have been generated with the BOEMIE shallow text processing technology.

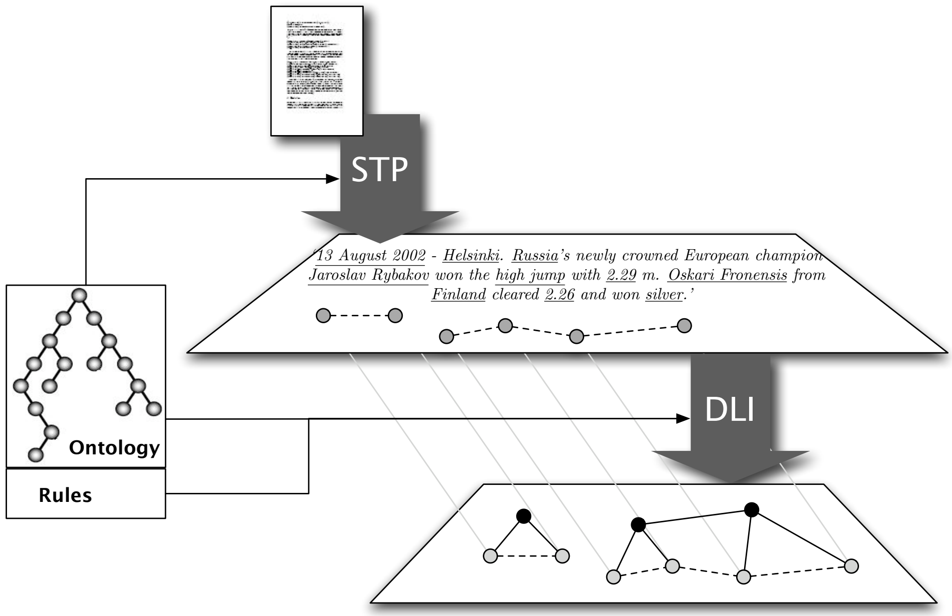

Fig. 1. Ontology based text interpretation through the use of shallow text processing (STP) results as input for deep-level interpretation (DLI).

- A two phase process of information extraction (IE), where the first phase exploits state-of-the-art shallow text processing mechanisms to extract surface-level structures as input for the second phase. The second phase called deep-level interpretation, exploits reasoning techniques over ontologies to extract deep-level semantics.
- The use of reasoning services, with abduction [1] as the key reasoning service for text interpretation.
- The use of ontologies, which provide the necessary index structures to represent surface-level, as well as deep-level semantics and to support logic-based IR.
- The use of ontologies as valuable and scalable assets, due to the use of Description Logics (DLs) [2] as their formal basis to support well studied reasoning tasks.
- The use of results from shallow text processing techniques, to extract surface-level semantics, are good enough to extract deep-level semantics and saves overhead to cope with large-scale corpora.

The previous hypotheses are the result of our work on the implementation and evaluation of a generic architecture for multimedia interpretation, which is described and evaluated here in the context of text interpretation. The approach followed for the design of this generic architecture is based on the combination of the works in [3], [4] and [5]. Different from [3] that employs processes for syntactic parsing, we argue that the results of shallow processing are good enough as input for reasoning techniques to extract deep-level semantics and to be able to deal with large-scale corpora. The evaluation results presented in this work confirm this hypothesis. In contrast to our work, in which we use DLs as knowledge representation formalism, [3] and [4] use first-order logics which have more expressive capabilities, but lack of automatic mechanism to prove for coherence. This is an important characteristic of DLs, therefore making this generic architecture a scalable one.

This work will focus on the most relevant part of the generic architecture, namely on the second phase of extraction where we show how the extraction of deep-level structures through abductive reasoning is useful for text interpretation. Section 2 introduces preliminaries for abduction as a core reasoning service for deep-level interpretation and presents its formalization in the context of DLs. Section 3 provides a real-world example to illustrate the interpretation process. Section 4 provides an empirical evaluation of the interpretation results generated by the architecture over a collection of web pages. Finally we conclude our work in Section 5

2 Abduction

Abduction is usually defined as a form of reasoning from effects to causes and aims at finding explanations (causes) for observations (effects). In this work text interpretation can be achieved through reasoning, more specifically through abduction. In general, abduction is formalized as

$$\Sigma \cup \Delta \models \Gamma \tag{1}$$

where background knowledge (Σ), and observations (Γ) are given and explanations (Δ) are to be computed.

DLs are a family of formal knowledge representation languages that support decidable reasoning problems [2]. If DLs are used as the underlying knowledge representation formalism, the background knowledge Σ is a knowledge base (KB) that consists of a Tbox $\mathcal{T}$ and an Abox $\mathcal{A}$: $\Sigma = (\mathcal{T}, \mathcal{A})$. In DL formalisms a KB consists of a Tbox that contains intentional knowledge in the form of a terminology and an Abox that contains the extensional (or assertional) knowledge that is specific to the individuals of the domain of discourse. Furthermore, Δ and Γ in Formula 1 are Aboxes and, therefore, they contain sets of role and concept instance assertions.

This work considers Abox abduction in DLs as the key inference service for text interpretation and the previous equation is modified to:

$$\Sigma \cup \Gamma_1 \cup \Delta \models \Gamma_2 \tag{2}$$

by splitting the assertions in Γ into two parts: bona fide assertions (Γ_1) and assertions requiring fiats (Γ_2). Bona fide assertions are assumed to be true by default, whereas fiat assertions are aimed to be explained.

In order to compute explanations, we use the implementation of Abox abduction as a non-standard retrieval inference service in DLs. Different from the standard retrieval inference services, answers to a given query cannot be found by simply exploiting the knowledge base. In fact, the abductive retrieval inference service has the task of acquiring what should be added to the knowledge base in order to positively answer a query.

To answer a given query, the abductive retrieval inference service can exploit non-recursive DL-safe rules with autoepistemic semantics [6] in a backward-chaining way. In this approach, rules are part of the knowledge base and are used to extend the expressivity of DLs. In order to extend expressivity and preserve decidability at the same time, the safety restriction is introduced for rules. Informally speaking, rules are DL-safe if they are only applied to Abox individuals, i.e., individuals explicitly named in the Abox [7]. In [8] a detailed discussion of the abductive retrieval inference service in DLs is presented.

The output of the abductive retrieval inference service should be a set Δs, which contains all explanations that are consistent w.r.t. Σ and Γ. However, in practice, one is not interested in retrieving every consistent explanation, but the most preferred explanation for every query. To achieve this goal, Δs is transformed into a poset according to a preference score. The preference score should reflect the two criteria proposed by Thagard for selecting explanations [9], namely simplicity and consilience. The less hypothesized assertions an explanation contains (simplicity) and the more fiat assertions (observations) an explanation involves (consilience), the higher its preference score should get. Therefore, the following formula to compute the preference score of each explanation is used: $S(\Delta) := S_f(\Delta) - S_h(\Delta)$. In this formula S_f is a term that reflects the involvement of fiat assertions in the explanation and S_h is a term that reflects the involvement of hypothesized assertions in the explanation. Thus, S_f and S_h can be defined as follows:

$$S_f(\Delta) := |\{i | i \in assertions(\Delta) \text{ and } i \in assertions(\Gamma_1)\}|$$
$$S_h(\Delta) := |\{i | i \in newAssertions(\Delta)\}|$$

The function *assertions* returns the set of all concept or role assertions found in a given Abox. The set *newAssertions* contains all concept or role assertions that are hypothesized during the generation of an explanation (hypothesized assertions).

3 An Example For Text Interpretation

In the context of a DL-based text interpretation architecture (like the one presented here) the results of shallow text processing can be represented as Aboxes. In order to extract deep-level semantics an Abox (hereafter called analysis Abox) is required as input. The input contains surface-level descriptions (see Figure 1) as results of shallow text processing. The interpretation process produces another Abox as output (interpretation Abox), which contains also deep-level besides surface-level semantic descriptions. The analysis Abox corresponds to Γ in the abduction formula (see Formula 1 in Section 2). The interpretation Abox is computed in an iterative process, and at the end of this process it contains the most preferred interpretation(s) of the text. The process starts with splitting Γ into bona fide (Γ_1) and fiat assertions (Γ_2). Afterwards, the interpretation process proceeds with the following tasks in each iteration:

First, each fiat assertion in Γ_2 is transformed into a corresponding query and sent to the abductive retrieval inference service. The abductive retrieval inference service returns the most preferred interpretation to answer this query. Second, the explanation is added to Γ_1. Furthermore, the fiat assertion which has been used to constitute the query is removed from Γ_2 and added to Γ_1. Third, Γ_1 is checked to find out whether new information can be inferred through deduction. If such information can be found it is added to Γ_1 as well.

At the end of each iteration, the interpretation process analyzes Γ_1 to identify new fiat assertions and starts a new iteration. In particular, assertions that are added during the previous iteration step are identified as fiat assertions for the new iteration. If no new fiat assertions can be identified, the interpretation process returns Γ_1 as the interpretation Abox and terminates.

In the following we present a step by step interpretation of a text to discuss the details of the interpretation process. Figure 2 shows a text excerpt from a web page with athletics news. The underlined words in Figure 2 are keywords of this text, which have to be detected by shallow text processing. The results of the first phase in the architecture for the text in Figure 2 are represented in an Abox (analysis Abox), which is shown in Figure 3.

'<u>13 August 2002</u> - <u>Helsinki</u>. <u>Russia</u>'s newly crowned European champion <u>Jaroslav Rybakov</u> won the <u>high jump</u> with <u>2.29</u> m. <u>Oskari Fronensis</u> from <u>Finland</u> cleared <u>2.26</u> and won <u>silver</u>.'

Fig. 2. Sample text paragraph with underlined tokens from shallow text processing.

To continue with the interpretation example, it is assumed that the ontology contains the axioms shown in Figure 4. In order to capture constraints among parts of aggregates, it is assumed that the ontology is extended with DL-safe rules (rules that are applied to Abox individuals only). In Figure 5 a set of rules for the athletics example is specified. Note that these rules define additional constraints on the concepts described in the ontology and, therefore, represent additional knowledge.

We assume that rules such as those shown in Figure 5 and the Tbox in Figure 4 constitute the background knowledge Σ. For the sake of brevity the Tbox and the set of rules show only a small excerpt of the athletics ontology, which is relevant for the text interpretation example discussed here.

To construct an interpretation for a text, explanations are searched to reveal why some words are related with some other words. Such explanations are then used to construct interpretation(s). Abox abduction (as presented in Section 2)

$$
\begin{aligned}
date_1 &: Date\\
(date_1, \text{'13 August 2002'}) &: hasValue\\
country_1 &: Country\\
(country_1, \text{'Russia'}) &: hasValue\\
hjName_1 &: HighJumpName\\
(hjName_1, \text{'high jump'}) &: hasValue\\
perf_2 &: Performance\\
(perf_2, \text{'2.26'}) &: hasValue\\
country_2 &: Country\\
(country_2, \text{'Finland'}) &: hasValue\\
(hjName_1,\ date_1) &: sportsNameToDate\\
(hjName_1,\ city_1) &: sportsNameToCity\\
(pName_1,\ country_1) &: personNameToCountry\\
(pName_2,\ country_2) &: personNameToCountry\\
city_1 &: City\\
(city_1, \text{'Helsinki'}) &: hasValue\\
pName_1 &: PersonName\\
(pName_1, \text{'Jaroslav Rybakov'}) &: hasValue\\
perf_1 &: Performance\\
(perf_1, \text{'2.29'}) &: hasValue\\
pName_2 &: PersonName\\
(pName_2, \text{'Oskari Fronensis'}) &: hasValue\\
rank_1 &: Ranking\\
(rank_1,\ \text{'}silver') &: hasValue\\
(pName_1,\ perf_1) &: personNameToPerformance\\
(pName_2,\ perf_2) &: personNameToPerformance\\
(hjName_1,\ perf_1) &: sportsNameToPerformance
\end{aligned}
$$

Fig. 3. Abox representing the results of shallow text processing.

is exploited to generate explanations and, therefore, constitutes the foundation of text interpretation in this architecture.

To start with the interpretation of the text paragraph in Figure 2, the shallow processing results for this text paragraph, namely the Abox in Figure 3, are considered as Γ. The following Formula 2 Γ is divided into a part Γ_2 that the agent would like to have explained (fiat assertions), and a part Γ_1 that the interpretation agent takes for granted (bona fide assertions). In this example Γ_2 is:

$$
\begin{aligned}
&(hjName_1,\ date_1) : sportsNameToDate,\\
&(pName_1,\ perf_1) : personNameToPerformance,\\
&(hjName_1,\ city_1) : sportsNameToCity,\\
&(pName_2, perf_2) : personNameToPerformance,\\
&(pName_1,\ country_1) : personNameToCountry,\\
&(hjName_1,\ perf_1) : sportsNameToPerformance,\\
&(pName_2,\ country_2) : personNameToCountry.
\end{aligned}
$$

$$
\begin{aligned}
Person \sqsubseteq\ & \exists hasName.PersonName \\
& \sqcap \exists hasNationality.Country \\
Athlete \sqsubseteq\ & Person \\
HighJumper \sqsubseteq\ & Athlete \\
PoleVaulter \sqsubseteq\ & Athlete \\
HighJumpName \sqsubseteq\ & SportsName \sqcap \neg PoleVaultName \\
PoleVaultName \sqsubseteq\ & SportsName \\
SportsTrial \sqsubseteq\ & \exists hasParticipant.Athlete \\
& \sqcap \exists hasPerformance.Performance \\
& \sqcap \exists hasRanking.Ranking \\
HighJump \sqsubseteq\ & SportsTrial \sqcap \forall hasParticipant.HighJumper \\
& \sqcap \neg PoleVault \\
PoleVault \sqsubseteq\ & SportsTrial \sqcap \forall hasParticipant.PoleVaulter \\
SportsRound \sqsubseteq\ & \exists hasName.RoundName \sqcap \exists hasDate.Date \\
& \sqcap \exists hasPart.SportsTrial \\
HighJumpRound \sqsubseteq\ & SportsRound \sqcap \forall hasPart.HighJump \\
& \sqcap \neg PoleVaultRound \\
PoleVaultRound \sqsubseteq\ & SportsRound \sqcap \forall hasPart.PoleVault \\
SportsCompetition \sqsubseteq\ & \exists hasPart.SportsRound \\
& \sqcap \exists hasName.SportsName \\
& \sqcap \exists takesPlace.City \\
HighJumpCompetition \sqsubseteq\ & SportsCompetition \\
& \sqcap \forall hasPart.HighJumpRound \\
& \sqcap \forall hasName.HighJumpName \\
& \sqcap \neg PoleVaultCompetition \\
PoleVaultCompetition \sqsubseteq\ & SportsCompetition \\
& \sqcap \forall hasPart.PoleVaultRound \\
& \sqcap \forall hasName.PoleVaultName
\end{aligned}
$$

Fig. 4. A tiny example TBox Σ for the athletics domain.

In the first step, these assertions are transformed into corresponding queries and the abductive retrieval inference service is asked for explanations. For example, from the role assertion $(hjName_1,\ date_1) : sportsNameToDate$ the following query is derived:

$Q := \{() \mid sportsNameToDate(hjName_1,\ date_1)\}$

The abductive retrieval inference service has the task of computing what should be added to the KB in order to answer this query with true. In the given set of rules (see Figure 5), there are two rules that have the atom $sportsNameToDate$ in the rule head (consequences). Both rules are applied in a backwardchaining way (i.e., from left to right) and corresponding terms are unified and variable bindings are obtained for X and Y. The unbound variable Z is instantiated with a new individual (e.g., new_ind_1). Note that for one of these rules, namely for the one that hypothesizes a pole vault competition, all bindings that are found for Y produce explanations that are inconsistent w.r.t. Σ. This is caused by the disjointness expressed in some of the concept description axioms in

$$
\begin{aligned}
personNameToCountry\,(X,Y) \leftarrow\ & Person(Z), \\
& hasPersonName(Z,X), \\
& PersonName(X), \\
& hasNationality(Z,Y), Country(Y). \\
personToPerformance\,(X,Y) \leftarrow\ & Person(X), \\
& hasPersonName(X,Z), \\
& PersonName(Z), \\
& personNameToPerformance(Z,Y). \\
personToPerformance\,(X,Y) \leftarrow\ & SportsTrial(Z), \\
& hasParticipant(Z,X), Athlete(X), \\
& hasPerformance(Z,Y), \\
& Performance(Y). \\
sportsNameToCity\,(X,Y) \leftarrow\ & HighJumpCompetition(Z), \\
& hasSportsName(Z,X), \\
& HighJumpName(X), \\
& takesPlace(Z,Y), City(Y). \\
sportsNameToCity\,(X,Y) \leftarrow\ & PoleVaultCompetition(Z), \\
& hasSportsName(Z,X), \\
& PoleVaultName(X), \\
& takesPlace(Z,Y), City(Y). \\
sportsNameToDate\,(X,Y) \leftarrow\ & HighJumpCompetition(Z), \\
& hasSportsName(Z,X), \\
& HighJumpName(X), \\
& hasDate(Z,Y), Date(Y). \\
sportsNameToDate\,(X,Y) \leftarrow\ & PoleVaultCompetition(Z), \\
& hasSportsName(Z,X), \\
& PoleVaultName(X), \\
& hasDate(Z,Y), Date(Y). \\
sportsCompetitionToPerformance\,(X,Y) \leftarrow\ & SportsCompetition(X), \\
& hasSportsName(X,Z), \\
& SportsName(Z), \\
& sportsNameToPerformance(Z,Y). \\
sportsCompetitionToPerformance\,(X,Y) \leftarrow\ & HighJumpCompetition(X), \\
& hasPart(X,Z), \\
& HighJumpRound(Z), \\
& hasPart(Z,W), HighJump(W) \\
& hasPerformance(W,Y). \\
sportsCompetitionToPerformance\,(X,Y) \leftarrow\ & PoleVaultCompetition(X), \\
& hasPart(X,Z), \\
& PoleVaultRound(Z), \\
& hasPart(Z,W), PoleVault(W) \\
& hasPerformance(W,Y).
\end{aligned}
$$

Fig. 5. Additional restrictions for text interpretation in the form of rules.

the TBox (e.g., the concepts *HighJumpName* and *PoleVaultName* are disjoint). The abductive retrieval service discards inconsistent explanations. Therefore, the explanation generated in order to answer Q with true is:

$\Delta_1 = \{new\_ind_1 : HighJumpCompetition, (new\_ind_1,\ date_1) : hasDate,$
$(new\_ind_1,\ hjName_1) : hasSportsName\}$

The assertions shown in Δ_1 are added to Γ_1. Furthermore the assertion $(hjName_1,\ date_1) : sportsNameToDate$ is removed from Γ_2 and added to Γ_1. This procedure is applied to the remaining assertions in Γ_2 until Γ_2 is empty. At the end of the first interpretation step, Γ_1 contains (beside the assertions shown in Figure 3) the following newly created assertions:

$new\_ind_1 : HighJumpCompetition,\ (new\_ind_1,\ hjName_1) : hasSportsName,$
$(new\_ind_1,\ date_1) : hasDate, (new\_ind_1,\ city_1) : takePlace, new\_ind_2 : Person,$
$(new\_ind_2,\ pName_1) : hasPersonName, (new\_ind_2,\ country_1) : hasNationality,$
$new\_ind_3 : Person, (new\_ind_3,\ pName_2) : hasPersonName,$
$(new\_ind_3,\ country_2) : hasNationality$

Note that the preference score presented in Section 2 guarantees that explanations that involve less hypothesized individuals and more observations are preferred. This is why Γ_1 contains a single *HighJumpCompetition* instance at the end of the first interpretation step.

In the second step, the interpretation process applies the set of rules in a forward chaining way (from right to left) to check whether new information can be deduced. This yields the following assertions:

$(new\_ind_2,\ perf_1) : personToPerformance,$
$(new\_ind_3,\ perf_2) : personToPerformance,$
$(new\_ind_1,\ perf_1) : sportsCompetitionToPerformance$

which are also added to Γ_1. At this state, the interpretation process defines a new Γ_2 by selecting all newly inferred assertions as fiat assertions and starts a new iteration. The first interpretation step is applied to the assertions in the new Γ_2. At the end of this step, the following newly created assertions are added to Γ_1:

$new\_ind_4 : HighJumpRound, (new\_ind_1,\ new\_ind_4) : hasPart,$
$new\_ind_5 : HighJump, (new\_ind_4,\ new\_ind_5) : hasPart,$
$(new\_ind_5,\ perf_1) : hasPerformance, (new\_ind_5,\ new\_ind_2) : hasParticipant,$
$new\_ind_6 : SportsTrial, (new\_ind_6,\ new\_ind_3) : hasParticipant,$
$(new\_ind_6,\ perf_2) : hasPerformance$

In the second step of the second iteration no new information can be deduced by applying the set of rules in a forward chaining way. Therefore, the interpretation process terminates by returning the current Γ_1 as the interpretation Abox. Besides the assertions in Figure 3, the interpretation Abox contains also the following newly inferred assertions:

$new\_ind_1 : HighJumpCompetition, new\_ind_2 : Person, new\_ind_3 : Person,$
$new\_ind_4 : HighJumpRound, new\_ind_5 : HighJump, new\_ind_6 : SportsTrial,$
$(new\_ind_1,\ hjName_1) : hasSportsName, (new\_ind_1,\ date_1) : hasDate,$
$(new\_ind_1,\ city_1) : takePlace, (new\_ind_1,\ new\_ind_4) : hasPart,$

$(new\_ind_4,\ new\_ind_5) : hasPart, (new\_ind_5,\ perf_1) : hasPerformance,$
$(new\_ind_5,\ new\_ind_2) : hasParticipant,$
$(new\_ind_6,\ new\_ind_3) : hasParticipant, (new\_ind_6,\ perf_2) : hasPerformance$
$(new\_ind_2,\ pName_1) : hasPersonName, (new\_ind_2,\ country_1) : hasNationality,$
$(new\_ind_3,\ pName_2) : hasPersonName, (new\_ind_3,\ country_2) : hasNationality$

Note that in the interpretation Abox the person instance new_ind_2 participates in a high jump trial (new_ind_5) and, therefore, is also an instance of the concept $HighJumper$ (see the Tbox in Figure 4). Thus, information about abstract events, e.g. high jump trials, also influences information that is available about the related parts. With queries for $HighJumper$s the corresponding text would not have been found otherwise. Thus, recognizing abstract events means extracting deep-level semantics that without reasoning is not possible to obtain from the surface of the text.

4 Evaluation

In this section, the utility of the architecture is analyzed through an empirical evaluation of its results over a collection of web pages. For this purpose, the text interpretation architecture described in Section 1 was implemented. The core component of this implementation is the DL-reasoner RacerPro in version 1-9-2 [10] that supports various inference services. The abductive retrieval inference service, which is the key inference service for text interpretation, is integrated into the DL-reasoner. The architecture gets analysis Aboxes, exploits various inference services, and returns preferred interpretation Aboxes as deep-level semantic descriptions.

To test the implementation, an ontology about the athletics domain was used, and two different corpora of web pages containing daily news about athletics events. Furthermore, extractors that implement shallow text processing and machine learning techniques were trained in order to obtain concept instances as well as relations between the instances (see Figure 3). The training process was performed with the help of an annotation tool over the first corpus of web pages. The annotation process is two-fold, in the first step annotators manually associate words in the text, with corresponding concepts in the ontology. For this purpose, concepts such as the following have been annotated, i.e. *PersonName*, *Country*, *City*, *Age*, *Gender*, *Performance*, *Ranking*, *SportsName*, *RoundName*, *Date* and *EventName*. Second, the annotated concepts are filled into relational tables corresponding to *Athletes*, *SportTrials*, *Rounds* and *Events*, in order to train the extractors for the extraction of relations between concept instances. After finalizing the training process, the second corpus has been analyzed automatically to detect concept instances and relations between them, such that for each web page in the second corpus an analysis Abox with corresponding assertions has been generated without manual annotation effort.

As discussed in Section 3, given a set of fiat assertions (relations between concept instances), the interpretation process aims to extract deep-level semantic descriptions in the form of Abox structures. Therefore the criteria used for the evaluation is to prove that for every fiat assertion, the expected deep-level descriptions are generated. To set up the evaluation, a set of boolean queries is defined, such that for each fiat assertion a corresponding query is found in the set and is executed. If the query is answered with true, then the expected deep-level structures for the corresponding fiat assertions were correctly extracted. For example, given the fiat assertion $(pName_1,\ country_1) : personNameToCountry$ it is expected that index structures (deep-level) for a person with $pName_1$ as name and $country_1$ as nationality exists, therefore the following query is defined:

$$Q_1 := \{() \mid Person(?x), hasPersonName(?x,\ pName_1),\\ hasNationality(?x,\ country_1)\}$$

For this evaluation a total of 85 web pages about athletics news were analyzed through shallow text processing. The results of shallow text processing was automatically analized to count the number of observations. According to the type of observation a set o queries was produced and executed in order to probe for the extraction of deep-level structures against the number of expected ones. The results of this evaluation can be observed in Figure 6.

Deep-level structures	**Expected**	**Extracted**
Person	48	48
SportsRound	10	10
Competition	99	99
SportEvent	189	189
SportsTrial with participant	326	326

Fig. 6. Results of deep-level interpretation.

In this way the evaluation results, pointed out that all expected abstract concepts (explanations) and their relations were extracted as long as shallow text processing could deliver relations between surface-level instances and the neccesary rules for interpretation (abducibles) exist.

5 Conclusion

As observed in Section 3, ontologies are useful means to provide index structures in order to represent deep-level semantics. Furthermore, deep-level semantics is represented as relational structures representing the contents of the text, which can not be directly extracted from its surface (readable part of the text). The empirical evaluation of this work indicates that good ontology design is crucial for the architecture to extract the expected deep-level semantic structures. Thus, ontology designers should invest in producing a coherent ontology and a set of rules that define the space of abduceable predicates (explanations). While most of the work should be invested in the correct design of the ontology, in the long term, it means a good return of investment due to the support for ontology concistency check provided by existing well studied reasoning mechanisms. We believe that in the near future well designed ontologies will become a highly valuable asset to enhance the precision and recall of information retrieval. Furthermore, the use of ontologies provides a generic architecture to interpret different types of text. Ontologies can be tailored towards any domain of interest, depending on the targeted text documents. The results of shallow text processing techniques are good enough to extract deep-level semantics for dealing with large-scale corpora, however it is not discarded that results from other linguistic techniques, i.e. syntactic analysis, can improve the results of deep-level interpretation. Finally, as it can be observed in Figure 1, only the first phase of extraction is media (in this case text) dependent, while the second phase is media independent, therefore applicable for the interpretation of other types of multimedia, e.g. for images promising results were presented in [11].

References

1. Elsenbroich, C., Kutz, O., Sattler, U.: A Case for Abductive Reasoning over Ontologies. In: Proc. OWL-2006: OWL Experiences and Directions Workshop. (2006)
2. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F., eds.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press (2003)
3. Hobbs, J.R., Stickel, M., Appelt, D., Martin, P.: Interpretation as abduction. Artificial Intelligence Journal **Vol. 63** (1993)
4. Shanahan, M.: Perception as Abduction: Turning Sensor Data Into Meaningful Representation. Cognitive Science **29** (2005) 103–134
5. Möller, R., Neumann, B.: On Scene Interpretation with Description Logics. In Christensen, H., Nagel, H.H., eds.: Cognitive Vision Systems: Sampling the Spectrum of Approaches. Number 3948 in LNCS. Springer (2006) 247–278
6. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. 2nd edition edn. Prentice-Hall, Englewood Cliffs, NJ (2003)
7. Möller, R., Neumann, B.: Ontology-based Reasoning Techniques for Multimedia Interpretation and Retrieval. In: Semantic Multimedia and Ontologies : Theory and Applications. (2007) To appear.
8. Espinosa, S., Kaya, A., Melzer, S., Möller, R., Wessel, M.: Multimedia Interpretation as Abduction. In: Proc. DL-2007: International Workshop on Description Logics. (2007)
9. Thagard, R.P.: The best explanation: Criteria for theory choice. The Journal of Philosophy (1978)
10. Haarslev, V., Möller, R., Wessel, M.: RacerPro User's Guide and Reference Manual Version 1.9.2 (2007)
11. Peraldi, S.E., Kaya, A., Melzer, S., Möller, R., Wessel, M.: Towards a media interpretation framework for the semantic web. The 2007 IEEE/WIC/ACM International Conference on Web Intelligence (WI'07) (2007)