Undefined 1 (2016) 1–5
IOS Press

From hyperlinks to Semantic Web properties using Open Knowledge Extraction

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Valentina Presutti [a,*], Andrea Giovanni Nuzzolese [a] and Sergio Consoli [a] and Aldo Gangemi [a,b] and Diego Reforgiato Recupero [a]

[a] *STLab, Institute of Cognitive Sciences and Technologies, National Research Council, via San Martino della Battaglia 44, 00185, Roma, Italy E-mail: valentina.presutti@cnr.it*
[b] *LIPN, Université Paris 13 - Sorbonne Cité - CNRS*
E-mail: {andrea.nuzzolese, sergio.consoli, diego.reforgiato}@istc.cnr.it, aldo.gangemi@cnr.it

Abstract. Open information extraction approaches are useful but insufficient alone for populating the Web with machine readable information as their results are not directly linkable to, and immediately reusable from, other Linked Data sources. This work proposes a novel paradigm, named Open Knowledge Extraction, and its implementation (Legalo) that performs unsupervised, open domain, and abstractive knowledge extraction from text for producing directly usable machine readable information. The implemented method is based on the hypothesis that hyperlinks (either created by humans or knowledge extraction tools) provide a pragmatic trace of semantic relations between two entities, and that such semantic relations, their subjects and objects, can be revealed by processing their linguistic traces (i.e. the sentences that embed the hyperlinks) and formalised as Semantic Web triples and ontology axioms. Experimental evaluations conducted on validated text extracted from Wikipedia pages, with the help of crowdsourcing, confirm this hypothesis showing high performances. A demo is available at `http://wit.istc.cnr.it/stlab-tools/legalo`.

Keywords: open knowledge extraction, open information extraction, abstractive summarisation, link semantics, relation extraction.

1. Populating the Semantic Web from natural language text

The vision of the Semantic Web is to populate the Web with machine understandable data so that intelligent agents will be able to automatically interpret its content - just like humans do by inspecting Web content - and assist users in performing a significant number of tasks, relieving them of cognitive overload.

The Linked Data movement [3] realised the first substantiation of this vision by bootstrapping the publication of machine understandable information, mainly taken from structured data (typically databases) or semi-structured data (e.g. Wikipedia infoboxes). However, a large part of the Web content consists of natural language text, hence a main challenge is to extract as much relevant knowledge as possible from this content, and publish it in the form of Semantic Web triples. This work aims to solve this problem by extracting relational knowledge that is "hidden" in hyperlinks, which can be either defined manually by humans (e.g. Wikipedia pagelinks) or created automatically by Knowledge Extraction (KE) systems (e.g. a KE system can automatically add links to Wikipedia pages or to local datasets of Semantic Web entities).

Current KE systems address the task of linking pieces of text to Semantic Web entities very well (e.g.

*Corresponding author. E-mail: valentina.presutti@cnr.it

owl:sameAs) by means of named entity linking methods, e.g. NERD[1] [41], FOX[2], conTEXT[3] [24], Dbpedia Spotlight[4], Stanbol[5], TAGME [14], Babelfy [29]. Some of them (e.g. NERD) also perform sense tagging, i.e. adding knowledge about entity types (rdf:type).

Nevertheless, it is desirable to enrich Web content with other semantic relations than owl:sameAs and rdf:type, i.e. factual relations between entities. A *pragmatic trace* of a factual relation between two entities is the presence of a hyperlink, which is associated with its *linguistic trace*, i.e. the text surrounding the hyperlink. In fact, when we include a link in a Web page, we usually have a semantic relation in mind between something we are referring within the page, i.e. subject, and something referred by the target page, i.e. object, and the text where the hyperlink is embedded often provides an explanation of what such relation is. For example, a link to "Usenet" in the Wikipedia page of "John McCarthy" suggests a semantic relation between those two entities, which is explained by the sentence: "McCarthy often commented on world affairs on the Usenet forums"[6].

Besides common sense, this hypothesis is also supported by a previous study [33], which describes the extraction of encyclopedic knowledge patterns for DBpedia types, based on links between Wikipedia pages. A user study showed that hyperlinks between Wikipedia pages determine relevant descriptive contexts for DBpedia entities at the type level, which suggests that these links mirror relevant semantic relations between entities.

A hyperlink in a Web page can be produced either by a human or a KE system (e.g., by linking a piece of text to a Wikipedia page, which in turn refers to a Semantic Web entity, i.e. a DBpedia entity). If a KE system recognises two or more entities in a sentence, there is a possibility that such sentence expresses some relation between them. For example, the following sentence:

> The New York Times reported that John McCarthy died. He invented the programming language LISP.

can be automatically enriched using a KE system by linking the text fragments "The New York Times", "John MacCarthy", and "LISP" to the Wikipedia pages wikipedia:The_New_York_Times[7], wikipedia:John_McCarthy_(computer_scientist) and wikipedia:Lisp_(programming_language) (respectively), resulting in the following:

> The New York Times reported that John McCarthy died. He invented the programming language LISP.

In this example, the three hyperlinks identify entities that are relevantly related by factual relations: "John McCarthy" with "The New York Times", and "John McCarthy" with "LISP". By generalising this concept, any recognised named entity in a sentence (even if not associated with an existing Web URI) can be treated as a potential hyperlink target (e.g. to a local knowledge base). In the rest of the paper we use examples with entities that can be resolved to DBpedia, for the sake of simplicity. Revealing the semantics of hyperlinks (either defined by humans or KE systems) has a high potential impact on the amount of Web knowledge that can be published in machine readable form.

In the Semantic Web era, such factual relations should be expressed as RDF triples where subjects, objects, and predicates have a URI (except for literal objects and blank nodes), and predicates are formalised as RDF/OWL properties, in order to facilitate their reuse and alignment to existing vocabularies, and for example to annotate hyperlinks with RDFa, within HTML anchor tags.

While subjects and objects can be mostly directly resolved through existing public or local Semantic Web entities, predicates are to be defined by performing "paraphrasing", a summarisation task that abstracts over the text (when needed) in order to design labels that are as close as possible to what a human would design for a Linked Data vocabulary. In this respect, [40] distinguishes between extractive and abstractive summarisation approaches. Extractive methods select pieces of texts from the original source in order to define a summary (i.e. they rely only on the available text), while abstractive techniques ideally rely on modeling the text, and then combining it with other resources and language generation techniques for generating a summary. Abstractive methods are usually applied to large documents to the aim of producing a meaningful summary of their content.

This work proposes to apply the guiding principle of abstractive techniques to open information extrac-

[1] http://nerd.eurecom.fr
[2] http://aksw.org/Projects/FOX.html
[3] http://context.aksw.org/app/
[4] http://dbpedia-spotlight.github.com/demo
[5] http://stanbol.apache.org
[6] Cf. http://en.wikipedia.org/wiki/John_McCarthy_(computer_scientist)
[7] wikpedia: stands for http://en.wikipedia.org/wiki/

Subject	Predicate	Object	Approach
John Stigall	received	a Bachelor of arts	extractive
John Stigall	received	from the State University of New York at Cortland	extractive
dbpedia:John_Stigall	myprop:receive_academic_degree	dbpedia:Bachelor_of_arts	abstractive
dbpedia:John_Stigall	myprop:receive_academic_degree_from	dbpedia:State_University_of_New_York	abstractive

Table 1: Comparison between relations resulting from extractive and abstractive approaches for the sentence "John Stigall received a Bachelor of arts from the State University of New York at Cortland".

tion as a novel contribution. Open information extraction refers to an open domain and unsupervised extraction paradigm. Existing open information extraction approaches are mainly extractive, hence showing a complimentary nature to what we present in this paper. They mostly focus on breaking text in meaningful fragments for building resources of relational patterns (e.g. PATTY [30][8], Wisenet [28][9]), in some cases disambiguated on external semantic resources such as WordNet[10]. Others focus on extracting facts, which are represented as simplified strings between entities (e.g. Open Information Extraction (OIE) [26][11]) that are not given a Semantic Web identity.

Knowledge extraction for the Semantic Web should instead include an abstractive step, which exploits a formal semantic representation of text, and produces output that is compliant with Semantic Web principles and requirements. The method described in this paper demonstrates this novel approach, called *open knowledge extraction* (OKE). For example, given the sentence:

> John Stigall received a Bachelor of arts from the State University of New York at Cortland.

Table 1 compares the extracted relations resulting from an extractive approach (such as OIE [26][12]) - the first two rows - and from an abstractive approach - the last two rows. The abstractive results exemplify the expected result of a OKE system. The main difference is that with the abstractive approach, subjects and objects are identified as Semantic Web entities, the predicate is as close as possible to what a human would define for a Linked Data vocabulary by possibly using terms that are not mentioned in the original text. In addition to what Table 1 shows, the predicate would be formally defined in terms of OWL axioms and possibly aligned with existing Semantic Web vocabularies.

1.1. Contribution

The main contributions of this work are:

- the introduction of *Open Knowledge Extraction* (OKE), a paradigm based on unsupervised, open domain, and abstractive knowledge extraction from text for producing directly usable machine readable information;
- an implementation of OKE, named *Legalo* that given an English sentence produces a set of RDF triples representing relevant factual relations expressed in the sentence, the predicates of which are formally defined in terms of OWL axioms;
- an evaluation of Legalo performed on a corpus of validated sentences from Wikipedia pages that provide evidence of factual relations. The results have been evaluated with the help of crowdsourcing and the creation of a gold standard, all showing high values of precision, recall, and accuracy;
- a discussion highlighting the current limits of the approach and possible ways of improving it, and including an informal comparison of the proposed method with one of the main existing open information extraction tools.

Additionally, the paper includes a brief description of a specific implementation of OKE, specialised for extracting the semantics of Wikipedia pagelinks, which has been evaluated in [38] showing promising results.

The paper is structured as follows: Section 2 introduces a novel paradigm named Open Knowledge Extraction. Sections 3 and 4 describe the implementation of an OKE system, named *Legalo*, focusing on the method implemented and the pipeline of components, respectively. Legalo has been evaluated with the help of crowdsourcing, as described in Section 5. Section 6 discusses the limits of the method and possible

[8] https://d5gate.ag5.mpi-sb.mpg.de/pattyweb/
[9] http://lcl.uniroma1.it/wisenet/
[10] http://wordnet.princeton.edu/
[11] http://openie.cs.washington.edu/
[12] Notice that this is the output of OIE for this sentence

ways to improve it, and informally compares Legalo with Open Information Extraction (OIE) [26]. Section 7 discusses relevant research work and finally, Section 8 summarises the contribution of this work and indicates future works.

2. Introducing Open Knowledge Extraction

According to [2], an Open Information Extraction (OIE) system: "facilitates domain independent discovery of relations extracted from text and readily scales to the diversity and size of the Web corpus". In other words, OIE revolutionised the information extraction paradigm by introducing unsupervised learning, domain-independence of the extracted relations, and the ability to scale both on size and heterogeneity dimensions of the Web. The Open Knowledge Extraction (OKE) paradigm poses its focus on making the extracted relations directly usable in a Semantic Web context.

An Open Knowledge Extraction (OKE) system is expected to perform unsupervised, open domain, and web scale extraction and to additionally have the following capabilities:

Relation assessment To assess if a natural language sentence provides an evidence of a relevant relation between a given pair of entities, which may be identified by hyperlinks; *relevant* here means that there are enough explicit traces in the sentence to support the existence of a (conceptual) relation;

Label generation To generate a predicate for this relation, with a label that is as close as possible to what a human would define for a Linked Data vocabulary;

Property formalisation To formalise this relation as an OWL object property with TBox axioms (conceptual level), as well as to produce ABox axioms (factual level) using that property.

More formally:

Definition 1. (Relevant relation)
Let s be a natural language textual sentence embedding some hyperlinks, and (e_{subj}, e_{obj}) a pair of entities mentioned in s, where e_{subj} and e_{obj} are the target entities referred by two hyperlinks in s, $\varphi_s(e_{subj}, e_{obj})$ is a relevant relation between e_{subj} and e_{obj}, expressed in s, with e_{subj} being the subject of φ and e_{obj} being its object. $\Lambda \equiv \{\lambda_1, ..., \lambda_n\}$ is a set of Linked Data labels generated by humans for $\varphi_s(e_{subj}, e_{obj})$. Finally, $\lambda^{'}$ is a label generated by an OKE system for $\varphi_s(e_{subj}, e_{obj})$.

An OKE system is able to assess the existence of $\varphi_s(e_{subj}, e_{obj})$, to generate a label $\lambda^{'}$ equal or very similar to $\lambda_i \in \Lambda$, and to formalise it as a Semantic Web property. Notice that not all relations are binary, for example events have time and space indexing, there are relations that naturally take more than two arguments e.g. *Mary gave a book to John, as present for his birthday*. For this reason an OKE system has to take into account the n-ary nature of relations and cope with expressing them as triples, given the pragmatic constraint of Semantic Web standard languages. This impacts on the complexity to assess the existence of $\varphi_s(e_{subj}, e_{obj})$ and to generate an adequate label for it, especially when φ_s is the projection of a n-ary relation.

3. Legalo: an OKE implementation that generates Semantic Web properties from text

One of the main contributions of this paper is the implementation (and evaluation, cf. Section 5) of an OKE system, named *Legalo*[13].

The method implemented by Legalo is based on six main steps:

1. internal formal representation of the sentence (*abstractive step*);
2. assessment of the existence of a relevant relation between pairs of entities identified in s, according to the content of the sentence;
3. extraction of relevant terms for the predicate (*extractive step*);
4. generation of the predicate label (*abstractive step*);
5. formal definition of the predicate within the scope of its linguistic evidence and formal representation (*abstractive step*) ;
6. alignment (whenever possible) to existing Semantic Web properties.

3.1. Frame-based formal representation of a sentence

Legalo relies on a set of rules to be applied to a frame-based formal representation G of the sentence s (cf. Definition 2). G is a RDF graph designed following a

[13] A demo of Legalo is available at `http://wit.istc.cnr.it/stlab-tools/legalo/`

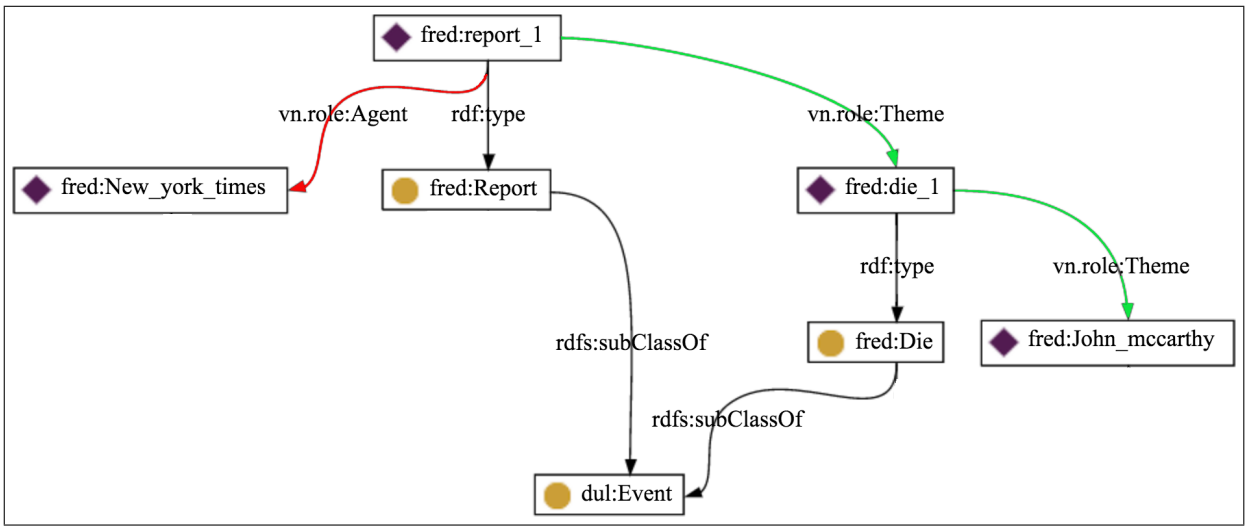

Fig. 1.: Frame-based formal representation for the sentence: "The New York Times reported that John McCarthy died."

frame-based approach, where nodes represent entities mentioned in s.

Definition 2. (Frame-based graph)
Let s be a natural language text sentence and $G = (V, E)$ a RDF (directed, multi-) graph modelling a frame-based formal representation of s, where $V \equiv \{v_0, ..., v_n\}$ is the set of nodes (i.e. subjects and objects from RDF triples) in G, $E \equiv \{edge_1, ..., edge_n\}$ is the set of edges (i.e. RDF triples) in G, where $edge_i = (v_{i-1}, p, v_i)$, is a triple connecting v_{i-1} and v_i with the RDF property p in G, and $v_i \in V$ is the node in G representing the entity e_i mentioned in s.

Frame Semantics [15] is a formal theory of meaning: its basic idea is that humans can better understand the meaning of a single word by knowing the relational knowledge associated to that word. For example, the sense of the word *buy* can be clarified in a certain context or task by knowing about the situation of a commercial transfer that involves certain individuals playing specific roles, e.g. a seller, a buyer, goods, money, etc.

In this work, frames are usually expressed by verbs or other linguistic constructions, and their occurrences in a sentence are represented as RDF n-ary relations, all being instances of some type of event or situation (e.g. `myont:buy_1 rdf:type myont:Buy`), which is on its turn represented as a subclass of `dul:Event`[14]. Intuitively, `dul:Event` is the top category of all frames expressed by verbs. In the context of this paper, the terms frame occurrence and event occurrences are used as synonyms. Entities that are mentioned in s are represented as individuals or classes, depending on their nature, which (ideally) have a type, defined based on the information available in the sentence s. When appropriate, entities are represented as arguments of n-ary relations, according to the role they play in the corresponding frame occurrence. The role of an entity in an event occurrence can be expressed either by a preposition, e.g. *Rico Lebrun taught **at** the Chouinard Art Institute*, or it can be abstracted from the text and represented by reusing the set of thematic roles defined by VerbNet [42], e.g. Rico Lebrun is the **agent** of the event occurrence "teach" in the above sample sentence.

A formal and detailed discussion of the theory behind frame-based formal representation of knowledge extracted from text, and used by Legalo is beyond the scope of this paper. This modeling approach and its founding theories are extensively described in [15,39, 32]. However, an example may be useful to convey the intuition behind the theory. Figure 1 shows a frame-based representation of the sentence:

> The New York Times reported that John McCarthy died.

[14] The prefix `dul:` stands for `http://www.ontologydesignpatterns.org/ont/dul/dul.owl\#`

The knowledge extracted from the sentence s is formalised as a set of RDF triples G. The figure is derived from the output of FRED [39] (see Section 4), the component providing the frame-based formal representation within Legalo. The prefix `fred:` stands for a local configurable namespace. Two entities can be identified in this sentence, i.e. "New York Times" and "John McCarthy", represented in G as individuals i.e. `fred:New_York_Times` and `fred:John_McCarthy`, respectively. Two frame occurrences can be identified in the sentence: one expressed by (an inflected form of) the verb *report* and the other expressed by the verb *die*. These frame occurrences are represented as n-ary relations: i.e., `fred:report_1` and `fred:die_1`, both being instances of classes (`fred:Report` and `fred:Die` respectively) that are of type `dul:Event`. Let us consider the event occurrence `fred:report_1`. Its arguments are: (i) `fred:New_York_Times`, which plays an *agentive role* in this event occurrence, formally expressed by the predicate `vn.role:Agent`[15], and `fred:John_McCarthy`, who plays a *passive role*, formalised by the predicate `vn.role:Theme`, both VerbNet thematic roles.

3.2. Relevant relation assessment

To assess if a relevant relation $\varphi_s(e_{subj}, e_{obj})$ exists in s between a pair of entities (e_{subj}, e_{obj}), Legalo relies on the analysis of the semantic structure of G. Firstly, $\varphi_s(e_{subj}, e_{obj})$ is assumed to hold only if there is at least one path in G connecting v_{subj} and v_{obj}, i.e. the nodes representing e_{subj} and e_{obj} in G, regardless of the edge direction in G. This is formally expressed by Axiom 1, given Definition 3.

Definition 3. (Graph path)
$G' = (V, E')$ is the undirected version of $G = (V, E)$. A path $P(v_{subj}, v_{obj}) = [v_0, edge_1, ..., edge_n, v_n]$ with $v_0 = v_{subj}$ and $v_n = v_{obj}$ is any sequence alternating nodes and edges in G' connecting v_{subj} to v_{obj}, or vice versa. The set $Pset_{subj,obj} \equiv \{edge_1, v_1, ..., edge_n\}$ includes all edges and nodes in $P(v_{subj}, v_{obj})$ excluding v_{subj} and v_{obj}.

Axiom 1. *(φ assessment: necessary condition)*

$$\varphi_s(e_{subj}, e_{obj}) \Rightarrow \exists P(v_{subj}, v_{obj})$$

If $P(v_{subj}, v_{obj})$ exists, Legalo distinguishes whether $P(v_{subj}, v_{obj})$ contains an event occurrence, or not. If $P(v_{subj}, v_{obj})$ does not contain any event occurrence, then the existence of $P(v_{subj}, v_{obj})$ is a sufficient condition to the existence of $\varphi_s(e_{subj}, e_{obj})$ (cf. Axiom 2).

Axiom 2. *(Assessment of φ: sufficient condition without event occurrences)*

$$\varphi_s(e_{subj}, e_{obj}) \Leftarrow \exists P(v_{subj}, v_{obj}) \; such \; that$$
$$\forall v_i \in Pset_{subj,obj}, \neg \texttt{dul:Event}(v_i)$$

In the other case, i.e. the path includes an event occurrence, $\varphi_s(e_{subj}, e_{obj})$ exists if e_{subj} is the subject of the event verb in the sentence. In the graph G this means that the node v_{subj} representing e_{subj} in G participates in the event occurrence with an agentive role. This is formalised by Axiom 3, given Definition 4.

Definition 4. (Agentive roles)
Let f be a node of G such that $\texttt{dul:Event}(f)$ $Role \equiv \{\rho_1, ..., \rho_n\}$ is the set of possible roles participating in f, $AgRole \equiv \{\rho_{m_1}, ..., \rho_{m_m}\}$ is the set of VerbNet agentive roles, with $AgRole \subseteq Role$, and $\rho(f, v_{subj})$ is a role connecting the event occurrence f to its participant v_{subj} (the node representing e_{subj}) in s.

Axiom 3. *(Assessment of φ with event occurrences: sufficient condition)*

$$\varphi_s(e_{subj}, e_{obj}) \Leftarrow \exists P(v_{subj}, v_{obj}) \; and$$
$$\exists f \in Pset_{subj,obj} \; such \; that \; \texttt{dul:Event}(f)$$
$$and \; \rho(f, v_{subj}) \in AgRole$$

This axiom is based on linguistic typology results (e.g. [8]), by which SVO (Subject-Verb-Object) languages such as English have almost always an explicit (or explicitatable) subject. This subject is formalized in a frame-based representation of s by means of an agentive role. Based on this observation, our method assumes that $\varphi_s(e_{subj}, e_{obj})$ exists if e_{subj} is the subject of a verb in s. This axiom is potentially restrictive with respect to the idea of a relevant relation expressed in a sentence, which may consider any pair of entities as related just because they are mentioned in a same sentence. In fact, this idea is quite difficult to implement, since relations between pairs of entities that play e.g. oblique roles (oblique roles are neither agentive or passive, e.g. "manner", "location", etc.) in a frame occurrence are hard to paraphrase even for a human. For example, consider the sentence:

[15] Prefix `vn.role:` stands for `http://www.ontologydesignpatterns.org/ont/vn/abox/role/`, which defines all VerbNet [42] thematic roles.

> After a move to Southern California in 1938, Rico Lebrun taught at the Chouinard Art Institute and then at the Disney Studios.

the frame-based representation of this sentence, depicted in Figure 2 identifies Rico Lebrun as the agent of a "teach" frame occurrence, while Southern California, Chouinard Art Institute, and Disney Studios participate in it with oblique roles. This sentence expresses three relevant relations: one between Rico Lebrun and Chouinard Art Institute, one between Rico Lebrun and Disney Studios, and another between Rico Lebrun and Southern California. All those relations can be summarised and represented as RDF triples, by Legalo.

While it is correct to state that Chouinard Art Institute and Disney Studios co-participate in an occurrence of the frame "teach", it is far from straightforward to paraphrase the meaning of this relation. E.g., one might say that Chouinard Art Institute and Disney Studios are both places where Rico Lebrun used to teach, but this paraphrase is not easily reconstructable from the text, and needs a stronger language generation approach, which has not been tackled for the moment. Additionally, such a paraphrase would not be usable for a binary predicate. A way to represent this relation is a generic co-participation relation, which is however too generic to be considered as relevant.

For this reason, the investigation of paraphrases of relation between entities co-participating in an event with oblique roles is left to further study. An interesting analysis on this problem that could suggest new work directions is discussed in [9].

3.3. *Combining extractive and abstractive design for property label generation*

As far as the generation of λ' is concerned (cf. Definition 1), Legalo combines extractive with abstractive techniques [40]. It means that it both reuses the terms in the text (extractive) and generates other terms derived from a semantic analysis of the text (abstractive). To this aim, it uses the semantic information provided by the frame-based representation G of the sentence s, which is further enriched with knowledge retrieved from external semantic resources. Legalo relies on the following knowledge resources:

- **DBpedia** [4] is the RDF version of Wikipedia and is used for resolving (disambiguating) the nodes $\{v_i\} \in V$ that represent the entities $\{e_i\}$ in the sentence s, on Linked Data;
- **Schema.org**[16] is a set of vocabularies for classifying entities on the Web. Schema.org is promoted by the most important search engines (Google, Yahoo!, Bing, and Yandex) making it a reference resource of its kind, which is why we decided to use it in Legalo for typing the recognised entities;
- **WiBi** [16] is a Wikipedia bitaxonomy: a refined, rich and high quality taxonomy that integrates Wikipedia pages and categories. Legalo uses WiBi as a reference semantic resource for designing the labels of generated properties, in particular for its "paraphrasing" task (i.e. abstractive step), when the extracted terms are too general to be informative enough; WiBi resulted the best resource to be used for this task as compared to the DBPedia ontology, and YAGO based on empirical tests conducted on a sample of $\sim$ 200 sentences;
- **VerbNet** [42] is the largest domain-independent hierarchical verb lexicon, available for English, which is one of the reasons why we decided to use it. Additionally, Legalo inherits it from FRED, its core component. VerbNet is organised into verb classes. Each verb class is described by thematic roles, selectional restrictions on the arguments, and frames. From VerbNet, Legalo obtains the thematic roles played by the DBpedia entities participating in a frame occurrence. Additionally, it is used for disambiguating the sense of frame occurrences. A subset of VerbNet thematic roles are mapped to specific prepositions, which are used in the paraphrasing task. The map is provided later in this section (cf. GR 3).

For example, consider the sentence:

> In February 2009 Evile began the pre-production process for their second album with Russ Russell.

Figure 3 shows the enriched frame-based formal representation of this sentence. The graph does not show WiBi types but they are actually retrieved by Legalo, for each resolved DBpedia entity. Two entities are resolved on DBpedia, i.e. `dbpedia:Evile`, and `dbpedia:Russ_Russell`, and two frame occurrences are identified, i.e. `fred:begin_1` and `fred:process_1`. Furthermore, each node is assigned with a type that, when possible, is aligned to existing Linked Data vocabularies. For Example,

[16] `http://schema.org/`

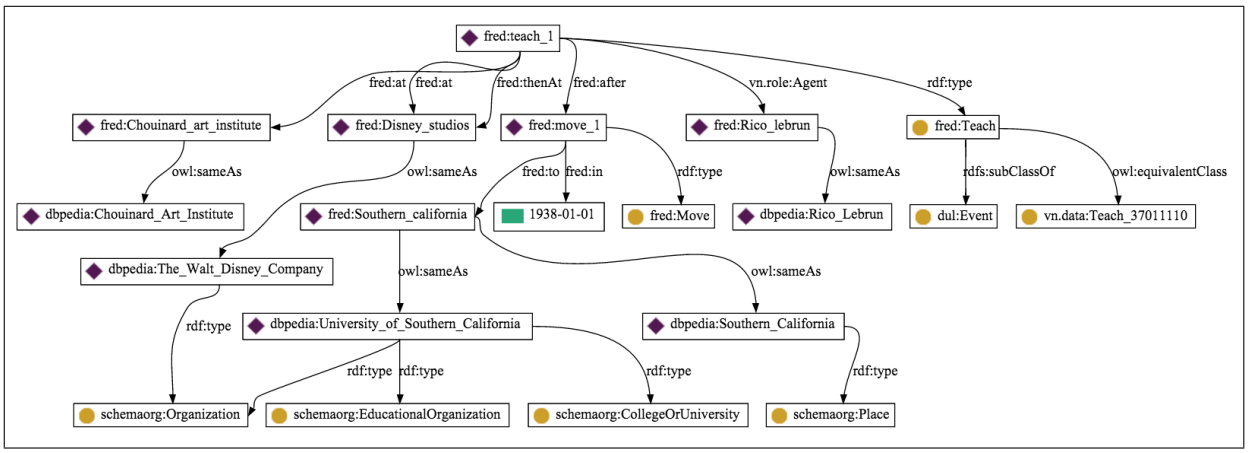

Fig. 2.: Frame-based formal representation for the sentence: "After a move to Southern California in 1938, Rico Lebrun taught at the Chouinard Art Institute and then at the Disney Studios". Legalo will select the pairs of entities (`fred:Rico_lebrun`, `Chouinard_art_institute`), (`fred:Rico_lebrun`, `fred:Disney_studios`), and (`fred:Chouinard_art_institute`, `fred:Southern_California`)

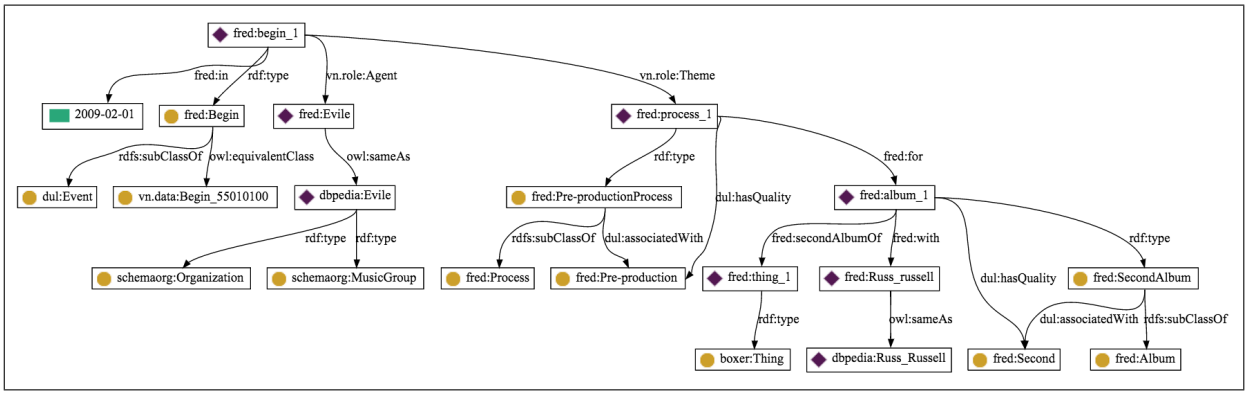

Fig. 3.: Frame-based formal representation for the sentence: "In February 2009 Evile began the pre-production process for their second album with Russ Russell" The graph is enriched with verb senses to disambiguate frame types, DBpedia entity resolutions, thematic roles played by DBpedia entities participating in frame occurrences, and entity types.

`dbpedia:Evile` has type `schema.org:MusicGroup` (Prefix `schema.org:` stands for `http://schema.org`), and the entity `fred:album_1` (representing the album mentioned in the sentence) is typed by the taxonomy `fred:SecondAlbum rdf:type fred:Album`. Following Axiom 1 and Axiom 3 (cf. Section 3.2), Legalo will select from the graph of Figure 3 the pair of (DBpedia) entities:
`dbpedia:Evile`, `dbpedia:Russ_Russell`
The Legalo design strategy for generating predicate labels is based on three main generative rules (GR). The first one concerns the concatenation of the labels that are used in the shortest path connecting the two nodes, including the labels of the edges and the labels of the node types in the path. This rule is defined by GR 1. It is important to remark that the path used as a reference for generating the predicate label is the one connecting the nodes v_{subj} and v_{obj} and not the corresponding resolved DBpedia entities.

GR 1. (Labels concatenation)
Given a pair (v_{subj}, v_{obj}):

- identify the shortest path(s) $P(v_{subj}, v_{obj})$ connecting v_{subj} and v_{obj};
- extract all labels (matching sentence terms) of the edges in the path;

– extract all labels of the most general types of the nodes that compose the path (if a node is typed by a taxonomy, the most general type in the taxonomy is extracted), except the types of v_{subj} and v_{obj};
– concatenate the extracted labels following their alternating sequence in $P(v_{subj}, v_{obj})$.

Hence, referring to Figure 3, Legalo will produce a predicate label λ = "begin process for album with" for expressing $\varphi_s(Evile, Russ\ Russell)$. Notice that the only labels that are included in the concatenation are those with prefix `fred:` meaning that they are extracted from s.

The second rule for generating predicate labels takes into account the possible presence of an event occurrence in the path connecting the pair (v_{subj}, v_{obj}). Intuitively, in this case the path is a tree, rooted in an event occurrence, i.e. a node f, such as `dul:Event`(f). The labels in this cases are extracted only from the path starting from f and ending in v_{obj} (referred as the right branch of the tree), including also the label of f type. The rationale behind this rule is that the right branch of the tree including the root event (i.e. its type) provides the relevant information expressing the relation between the two nodes, according to an empirical observation conducted on a sample of ∼200 cases.

For example, consider the (excerpt of the) frame-based representation of the sentence "Joey Foster Ellis has published on The New York Times, and The Wall Street Journal." shown in Example 3.1

Example 3.1. (Path including an event)

```
fred:publish_1 rdf:type fred:Publish;
  vn.role:Agent fred:Joey_Foster_Ellis;
  fred:on fred:New_York_Times;
  fred:on fred:Wall_Street_Journal .
fred:Publish rdfs:subClassOf dul:Event .
fred:Joey_Foster_Ellis
  owl:sameAs dbpedia:Joey_Foster_Ellis .
fred:New_York_Times
  owl:sameAs dbpedia:The_New_York_Times .
fred:Wall_Street_Journal
  owl:sameAs dbpedia:Wall_Street_Journal.
```

Following GR 1 and applying this additional rule for the selected pair:
`dbpedia:Joey_Foster_Ellis, dbpedia:Wall_Street_Journal`
leads to a predicate λ = "publish on" for $\varphi_s(Joey\ Foster\ Ellis, Wall\ Street\ Journal)$.

Additionally, if the right branch of the tree path is of length 1 and the only edge is a passive role, i.e. v_{obj} participates with a passive role to f, the label of the WiBi type of v_{obj} is concatenated to the predicate label. The rationale behind this rule is that when v_{subj} and v_{obj} play respectively an agentive and a passive role in an event occurrence, the resulting predicate label following only GR 1 would be too generic, hence adding the WiBi type label makes the property label more specific and informative.

For example, a frame-based representation of the sentence "Elton John plays the piano" is given in Example 3.2:

Example 3.2. (Right branch of tree path with only passive role)

```
fred:play_1 rdf:type fred:Play;
  vn.role:Agent fred:Elton_John;
  vn.role:Theme fred:piano_1 .
fred:Elton_John
  owl:sameAs dbpedia:Elton_John .
fred:piano_1 rdf:type dbpedia:Piano .
fred:Play rdfs:subClassOf dul:Event .
dbpedia:Piano
  rdf:type wibi:MusicalInstrument .
```

If we apply the additional rules described so far to the pair `(dbpedia:Elton_John, dbpedia:Piano)` we obtain a label λ = "play musical instrument" for $\varphi_s(Elton\ John, piano)$, which is more informative than a simple "play" that would result without adding the WiBi type label of `dbpedia:Piano`. This rule is defined by GR 2.

GR 2. (Path including event occurrences)
Given a selected pair (v_{subj}, v_{obj}) and the shortest path $P(v_{subj}, v_{obj})$ connecting them. If $P(v_{subj}, v_{obj})$ is a tree rooted in f, such as `dul:Event`(f),

– extract $P(f, v_{obj})$ from $P(v_{subj}, v_{obj})$;
– extract all edge labels in $P(v_{subj}, v_{obj})$ that match with terms extracted from s;
– for each v_i (including f and excluding v_{obj}) in $P(f, v_{obj})$ extract the label of its more general type;
– concatenate the extracted labels following their alternating sequence in $P(f, v_{obj})$;
– if $P(f, v_{obj})$ has only 1 edge (length = 1), and this edge identifies a VerbNet passive role, than extract the WiBi type of v_{obj} and append it to the label concatenation.

The third rule for predicate label generation complements GR 1 and GR 2 by associating VerbNet roles to labels. Such labels have been defined top-down by analysing VerbNet thematic roles and their usage examples. The rule is defined in GR 3.

GR 3. (Thematic roles labels)
If a path contains a VerbNet thematic role, replace its label with an empty one, unless the role is associated with a non empty label according to the following scheme:
`vn.role:Actor1` -> "with"
`vn.role:Actor2` -> "with"
`vn.role:Beneficiary` -> "for"
`vn.role:Instrument` -> "with"
`vn.role:Destination` -> "to"
`vn.role:Topic` -> "about"
`vn.role:Source` -> "from"

For example, consider the (excerpt of the) frame-based representation of the sentence "Lincoln's wife suspects that John Wilkes Booth and Andrew Johnson conspired to kill Lincoln." shown in Example 3.3.

Example 3.3. (Thematic roles associated with labels)

```
fred:conspire_1 rdf:type fred:Conspire;
  vn.role:Actor1 fred:Andrew_Johnson;
  vn.role:Actor2 fred:John_Wilkes_Booth .
fred:Conspire rdfs:subClassOf dul:Event .
fred:Andrew_Johnson
  owl:sameAs dbpedia:Andrew_Johnson .
fred:John_Wilkes_Booth
  owl:sameAs dbpedia:John_Wilkes_Booth;
fred:Lincoln
  owl:sameAs dbpedia:Abraham_Lincoln .
```

By applying GR 1, 2 and 3 to the path connecting the pair:
`dbpedia:Andrew_Johnson,`
`  dbpedia:John_Wilkes_Booth`
Legalo generates a label λ = "conspire with" for $\varphi_s(Andrew\ Johnson, John\ Wilkes\ Booth)$.
The mapping scheme (role<->label) is an evolving resource, which improves based on the periodic evaluation of Legalo outputs.

3.4. Formalisation of extracted knowledge

Given a textual sentence s and its frame-based formal representation G, by following the generative rules described in Section 3.3 Legalo generates a label λ for each relation $\varphi_s(e_{subj}, e_{obj})$ that it is able to identify in s, based on the shortest path $P(v_{subj}, v_{obj})$ connecting (v_{subj}, v_{obj}) in G (cf. Definitions 1, 2, and 3). These labels constitute the basis for automatically generating a set of RDF triples that can be used for semantically annotating the hyperlinks included in s. Additionally, these set of triples provides a (formalised) summary of s.

The aim of the formalisation step is to favour the reuse of the extracted knowledge by representing it as RDF triples, by augmenting it with informative annotations and axiomatisation, and by linking it to existing Semantic Web data. In particular, the formalisation step addresses the following tasks:

- producing a RDF triple $(v_{subj}, p_\lambda, v_{obj})$ for each hyperlink in s associated with e_{obj}, such that $\varphi_s(e_{subj}, e_{obj})$ exists in s, where p_λ is a predicate having label λ, v_{subj} is the node in G representing e_{subj}, and v_{obj} is the node in G representing e_{obj};
- formally defining p_λ: its domain and range, and possible other OWL axioms that specify its formal semantics;
- annotating each triple $(v_{subj}, p_\lambda, v_{obj})$ with information about its linguistic evidence, i.e. the sentence s;
- annotating each triple and predicate with information about the frame-based formal representation from which they were extracted.

RDF triples can be used for annotating hyperlinks, e.g. with RDFa, OWL axiomatisation supports ontology reuse, and scope annotations (i.e. linguistic evidence and formal representation) support reuse in relation extraction systems, e.g. relation extraction based on distant supervision [27,1]

Locality of produced predicates. Our method works on the assumption that each generated predicate and its associated formalisation are valid in the conceptual scope identified by the sentence s. This means that s identifies the scope of predicate names definitions, i.e. the namespace of a predicate depends on s. Pragmatically, this is implemented in Legalo by including the checksum of s in the predicate namespace. This strong locality constraint may lead to producing a high number of potentially equivalent properties (i.e. having the same intensional meaning) defined as they were different. This issue is tackled by formalising all predicates with domain and range axioms having values, i.e. classes, from external (open domain) resources, as

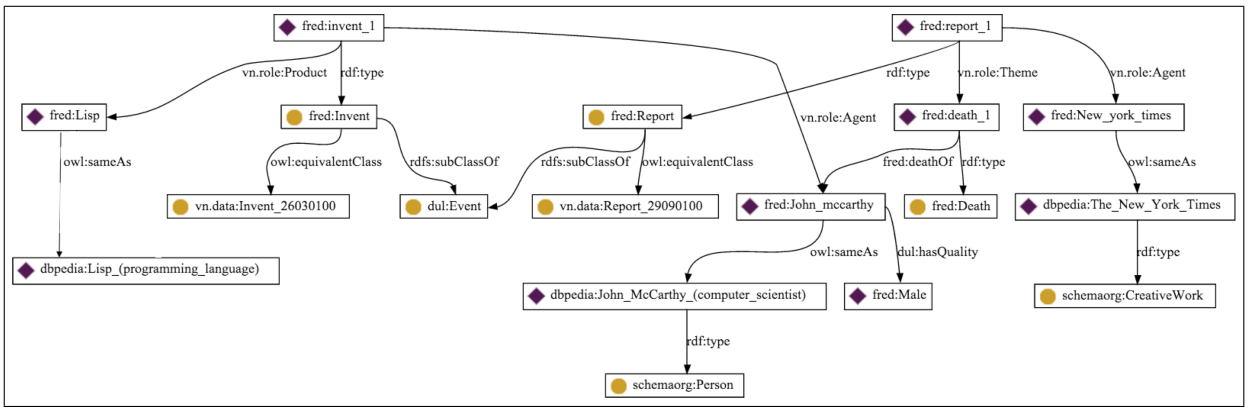

Fig. 4.: Frame-based formal representation for the sentence: "The New York Times reported the death of John McCarthy. He invented LISP."

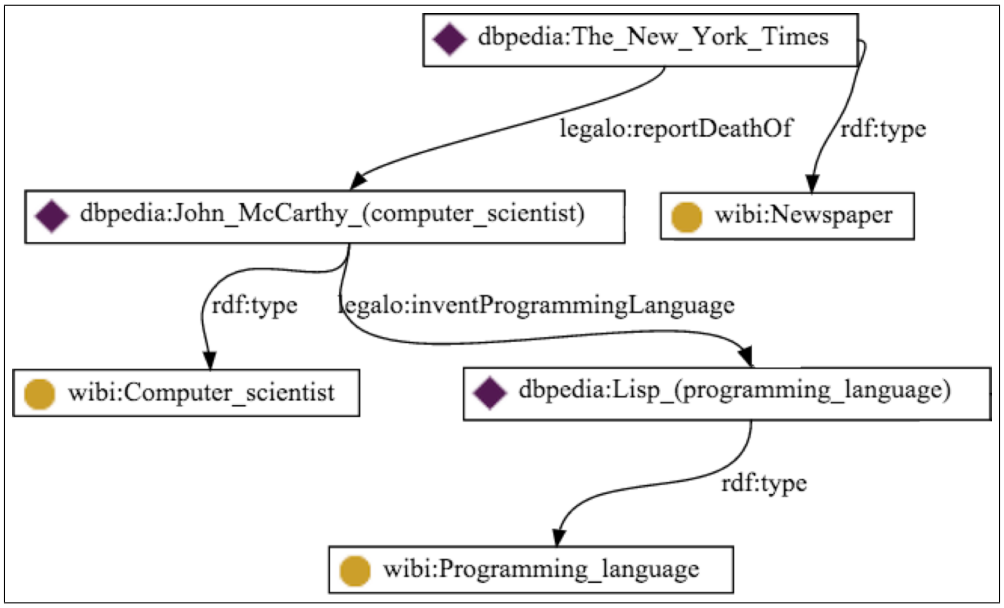

Fig. 5.: Legalo's triples produced from the sentence:"The New York Times reported the death of John McCarthy. He invented LISP. "

well as by keeping the binding between a predicate, its linguistic evidence, i.e. s, and its formal representation source, i.e. G. The latter contains information about the disambiguated senses of the verbs, i.e. frame occurrences, used in s. All these features allow on one hand to inspect a specific property for understanding its meaning, e.g. in case of manual reuse, on the other hand to automatically reconcile predicates by computing a similarity measure based on them. In this paper, we focus on the generative part of the problem, i.e. generating usable labels for predicates and producing their formal definition, while we leave the reconciliation task to future work.

RDF factual statements. For each hyperlink in s associated with a true assessment of $\varphi_s(e_{subj}, e_{obj})$ (cf. Axioms 1, 2, and 3), Legalo produces at least one RDF triple. As explained in Section 3.3, the nodes v_{subj} and v_{obj} in G representing e_{subj} and e_{obj} are resolved on DBpedia, when possible, which links the triples to the Linked Data cloud. The predicate is formalised as an OWL object property having λ' as label and an ID de-

rived by transforming λ' according to the CamelCase notation[17].

For example, consider the enriched frame-based formal representation of the sentence

> The New York Times reported the death of John McCarthy. He invented the programming language LISP.

depicted in Figure 4, Legalo produces the triples depicted in Figure 5, according to the generative rules GR 1, 2, and 3, where the prefix `legalo:` is a namespace defined using the checksum of the sentence s. Notice that Figure 5 shows the WiBi types[18] for the resolved DBpedia entities.

OWL property formalisation. For each generated property, Legalo produces an additional set of OWL axioms that formally define it. The predicate formalisation states that the predicate is an OWL object property, and includes domain and range axioms, whose values are defined according to the WiBi types assigned to v_{subj} and v_{obj}. In case of multi-typing of an entity, the value is the union of all types. In case a WiBi type is not available, the default type is `owl:Thing`. Example 3.4 shows the axioms formalising domain and range of the properties shown in Figure 5.

Example 3.4. (Domain and range axioms.)

```
legalo:reportDeathOf a owl:ObjectProperty
;
  rdfs:domain wibi:Newspaper ;
  rdfs:range wibi:Computer_scientist .
legalo:inventProgrammingLanguage
  a owl:ObjectProperty ;
  rdfs:domain wibi:Computer_scientist ;
  rdfs:range wibi:Programming_language ;
  rdfs:subPropertyOf legalo:invent .
```

As the reader may notice, an additional `rdfs:subPropertyOf` axiom is included in the formal definition of `legalo:inventProgrammingLanguage`. In fact, if a predicate is derived with GR 2, meaning that v_{subj} and v_{obj} participate in an event with respectively, an agentive and a passive role, then Legalo also generates a more general property based on the event type, and produces a `rdfs:subPropertyOf` axiom. We remind that in these cases, the rule requires to generate a specialised property label by appending the WiBi type of v_{obj} to the label of the event type. Example 3.4 shows one of this cases. All properties produced by Legalo are derived from a formal representation G of the sentence s, meaning that G provides their formal scope. Based on this principle, Legalo produces an additional set of triples, which formalise the generated properties with reference to G. As stated by GR 1 and 2, there are two main types of paths from which the properties can derive. In the first case, the path connecting v_{subj} and v_{obj} does not include any event node. In this case, Legalo produces a OWL property chain axiom stating that the generated property is implied by the chain of properties participating in the path, where each property of the path is formalised with domain and range axioms according to the locality of G. The same concept applies to the case of a path that includes an event node. Similarly, Legalo produces a property chain axiom. However, in this case the path has two different directions in G. For this types of paths we define the concepts of *left branch path*, i.e. the one connecting the event node with v_{subj}, and *right branch path*, i.e. the one connecting the event node with v_{obj}. For example, in Figure 4 the path P connecting `fred:John_Mccarthy` with `fred:Lisp` includes an event, i.e. `fred:invent_1`. Hence P is a tree, which root is this event node. The left branch path of P is the one connecting `fred:invent_1` with `fred:Lisp`, while the right branch path of P is the one connecting `fred:invent_1` with `fred:John_Mccarthy`. In order to define a property chain axiom Legalo needs to define the inverses of all properties in the left branch of P. However, these branch paths may contain properties defined by VerbNet, i.e. thematic roles, which are independent of the event they are associated with, in the scope of G., i.e. they are general domain properties. In other words, these properties do not carry any information about the event included in the path, which is relevant as far as the formal semantics of the generated property is concerned. Legalo tackles this issue by defining a local thematic role property for each VerbNet role participating in the event included in the path. For example, let us consider the property `legalo:inventProgrammingLanguage` in Figure 5. Its reference path includes the two (thematic roles) properties `vn.role:Agent` and `vn.role:Product`. Legalo generates two new properties, `legalo:AgentInvent` and `legalo:ProductInvent`, defined as sub-properties of `vn.role:Agent` and `vn.role:Product`, respec-

[17] According to a common Linked Data convention, using the CamelCase notation for OWL object properties makes the first term of the ID start with lower case, e.g. "invent programming language" -> inventProgrammingLanguage.

[18] `http://www.wibitaxonomy.org/`

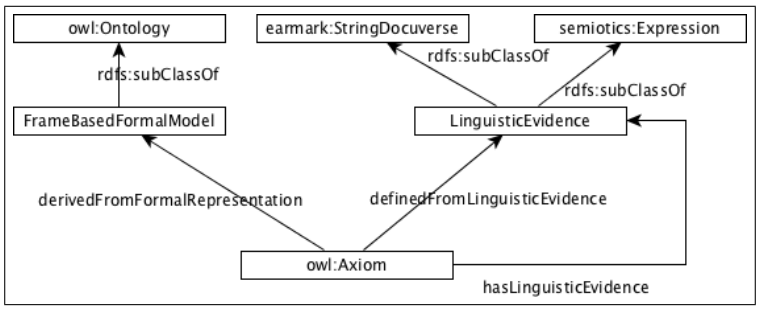

Fig. 6.: The grounding vocabulary used for annotating the generated triples and properties with information about their linguistic and formal representation scope.

tively. Given these two new properties, the axioms produced for formalising the generated property `legalo:inventProgrammingLanguage` are given in Example 3.5

Example 3.5. (Property Chain Axiom when the connecting path includes an event.)

```
legalo:inventProgrammingLanguage
  a owl:ObjectProperty ;
  owl:propertyChainAxiom _:b1 .
legalo:AgentInvent
  rdfs:subPropertyOf vn.role:Agent .
legalo:ProductInvent
  rdf:subPropertyOf vn.role:Product .
_:b1:-([owl:inverseOf legalo:AgentInvent]
  legalo:ProductInvent) .
```

Scope annotations. Finally, Legalo annotates all generated properties and triples with information related to the linguistic and formal representation scopes from which they were derived. To this aim a specific OWL ontology has been defined, named *grounding*[19], depicted in Figure 6. This ontology reuses Earmark[20], a vocabulary for annotating textual content, and *semiotics*[21], a content ontology pattern that encodes a basic semiotic theory. Earmark defines the class `earmark:Docuverse`, which represents any container of strings that may appear in a document. In the context of Legalo this class can be used for representing the sentence s. The semiotics content pattern defines three main classes: Expression, Meaning, Reference (the semiotic triangle). The class Expression is also reused for representing the sentence s. As for the annotation of the linguistic scope of a RDF triple, the grounding vocabulary defines the more specific concept of "linguistic evidence". In fact, according to the axioms defined in Section 3.2 and the generative rules defined in Section 3.3, the sentence s provides an evidence of the relation $\varphi_s(e_{subj}, e_{obj})$, which is formalised by a RDF triple $(v_{subj}, p_\lambda, v_{obj})$. The concept of "linguistic evidence" is represented by the class `LinguisticEvidence` that specialises both `earmark:Docuverse` and `semiotics:Expression`. The OWL property that relates a RDF triple generated by Legalo and its linguistic evidence is `hasLinguisticEvidence`.
Additionally, the class `FrameBasedFormalModel` is defined for representing the concept of *frame-based formal representation* of a textual sentence, described in detail in Section 3.1. This class is instantiated by the graph G representing s, which provides the formal scope for all generated properties and triples. The property `derivedFromFormalRepresentation` of the grounding ontology, connects a Legalo generated property as well as a RDF triple, with the graph G from which they were derived. As an example, let us consider the sentence represented by the graph in Figure 4 and the generated RDF triple of the property `legalo:inventProgrammingLanguage` depicted in Figure 5. The scope annotations shown in Example 3.6 are generated.

Example 3.6. (Scope annotations of a generated property)

```
legalo:sentence
  a grounding:LinguisticEvidence ;
  earmark:hasContent "The New York Times
reported the death of McCarthy. He invented
LISP." .
```

[19] The vocabulary can be downloaded from `http://ontologydesignpatterns.org/cp/owl/grounding.owl`

[20] `http://www.essepuntato.it/2008/12/earmark`

[21] `http://www.ontologydesignpatterns.org/cp/owl/semiotics.owl`, prefix `semio:`

```
[] a owl:Axiom ;
  grounding:hasLinguisticEvidence
    legalo:sentence;
  owl:annotatedProperty
    legalo:inventProgrammingLanguage ;
  owl:annotatedSource
    dbpedia:John_McCarthy_(computer_scientist);
  owl:annotatedTarget
    dbpedia:Lisp_(programming_language) .
legalo:inventProgrammingLanguage
  a owl:ObjectProperty ;
  grounding:derivedFromFormalRepresentation
    krgraph:52f88ca22 ;
  grounding:definedFromLinguisticEvidence
    legalo:sentence .
```

The first two axioms simply create an individual of type `LinguisticEvidence` for representing the sentence. The second group of axioms annotates the RDF triple for "John McCarthy invented Lisp" with its linguistic evidence.
Finally, the `legalo:inventProgrammingLanguage` property is annotated with its linguistic as well as its formal scope.

3.5. Alignment to Semantic Web vocabularies

This step has the goal of aligning the generated properties to existing Semantic Web ones. The idea is to maximise reuse and linking of extracted knowledge to existing Linked Data. Legalo implements a simple string matching technique based on the Levenshtein distance measure for addressing this task. The implementation of more sophisticated approaches for aligning generated properties to existing vocabularies is part of future work. Relevant related work are ontology matching techniques such as [13] (cf. see the Ontology Alignment Evaluation Initiative[22]). A possible strategy is to apply state-of-the-art techniques in ontology matching exploiting the information and features provided by the formalisation step (cf. Section 3.4).
Legalo uses three semantic resources for identifying possible targets for property alignment:

- **Watson**[23] [10] is a service that provides access to Semantic Web knowledge, in particular ontologies;
- **Linked Open Vocabularies (LOV)**[24] is an aggregator of Linked Open vocabularies (including DBpedia), and provides services for accessing their data;
- **Never-Ending Language Learning (NELL)**[25] [6] is a machine learning system that extracts structured data from unstructured Web pages and stores it in a knowledge base. It runs continuously since 2010. From the learnt facts, NELL team has derived an ontology of categories and properties: it includes 548 properties at the moment[26].

In principle other resources can be added and could be selected, we chose these three resources because they allow us to both cover most of public linked data vocabularies (i.e. LOV and Watson), and test with automatically generated resources (i.e. NELL).

4. Legalo pipeline and components

Legalo is based on a pipeline of components and data sources, executed in the sequence illustrated in Figure 7.

1. FRED: Semantic Web machine reader. The core component of the system is *FRED* [39], a Semantic Web machine reader able to produce a RDF/OWL frame-based representation of a text. It integrates the output of several NLP tools, enriches and transform it by reusing Linguistic Frames [32], Ontology Design Patterns [20], open data, and various vocabularies. FRED detects events, roles, and n-ary relations and represent in a RDF/OWL graph. It also represent variable discourse referents, such as the variable in the first-order predication *Cat(x)* extracted from the sentence *The cat is on the mat*, they are formalised as reified individuals e.g. `cat_1`. As far as Legalo is concerned, the most used features of FRED are the frame-based graph representation based on VerbNet verbs and thematic roles, the Named Entity Recognition (NER) and Resolution component i.e., TAGME [14], and the annotation of text fragments, based on the Earmark vocabulary and annotation method [35].

All figures depicted in Section 2 show examples of FRED outputs: the reader may want to consider Figure 3, which show the RDF/OWL graph for the sentence "In February 2009 Evile began the pre-production pro-

[22] http://oaei.ontologymatching.org/
[23] http://watson.kmi.open.ac.uk/WatsonWUI/
[24] http://lov.okfn.org/dataset/lov/
[25] http://rtw.ml.cmu.edu/rtw/
[26] http://nell-ld.telecom-st-etienne.fr/

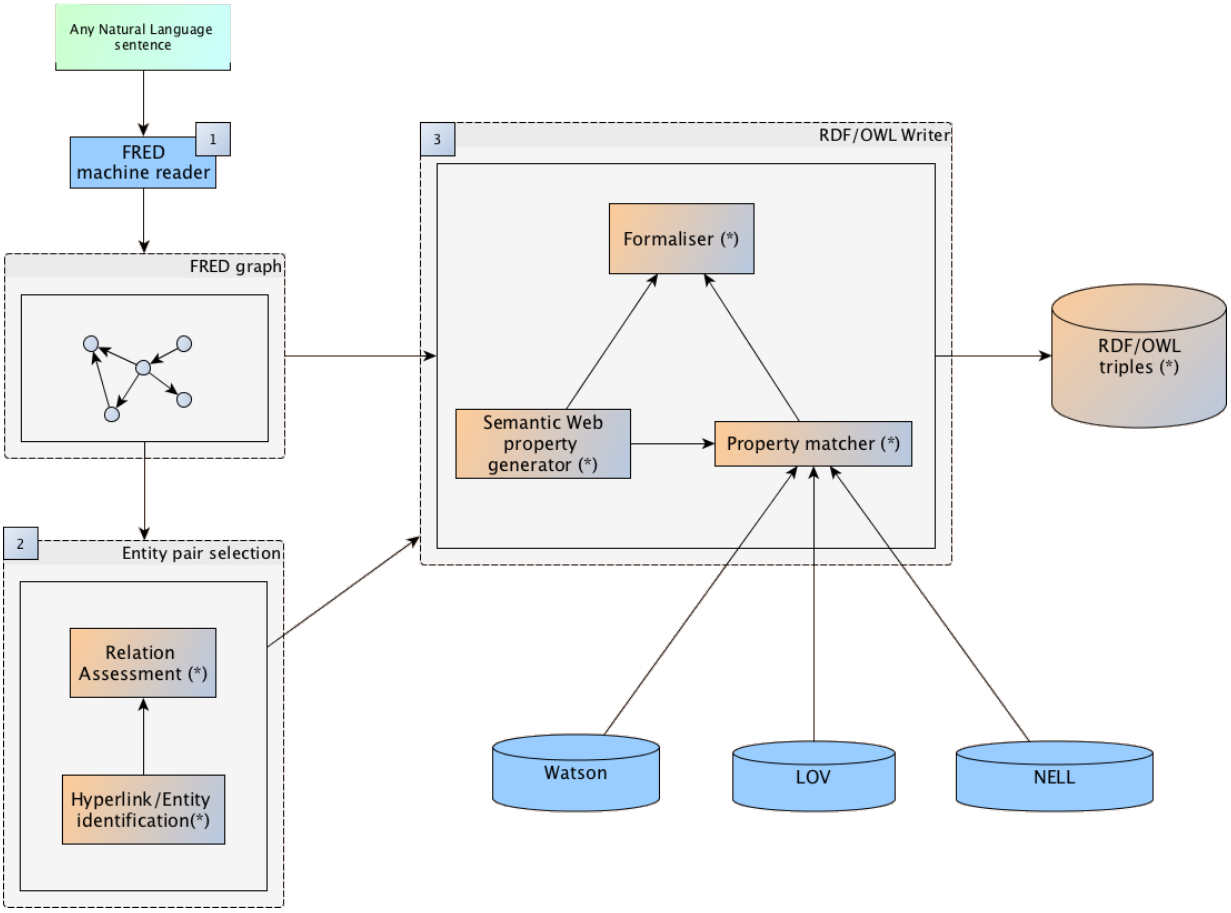

Fig. 7.: Pipeline implemented by Legalo for generating Semantic Web properties for semantic annotation of hyperlinks based on their linguistic trace, i.e. natural language sentence including the hyperlinks. Numbers indicate the order of execution of a component in the pipeline. Edges indicates input/output flows. (*) denotes tools developed in this work, which are part this paper contribution.

cess for their second album with Russ Russell" as a representative output of FRED.

2. Entity pair selection. This component is in charge of detecting the resolved entities and associate them with their lexical surface in s. This is done by querying FRED text span annotations. Another task of this component is, for each pair of detected entities (v_{subj}, v_{obj}), to assess the existence of φ_s between them. In other words, this component checks the existence paths between v_{subj} and v_{obj} (cf. Axiom 1), selects the shortest one and verifies if there are event nodes in the selected path. If so, it verifies if v_{subj} participates in the event occurrence with an agentive role (cf. Axiom 3). All selected pairs and associated paths are passed to the next component.

3. RDF/OWL writer. This component is in charge of generating a predicate for each pair of entities received in input from the previous component, by applying the generative rules described in Section 3.3 to its associated path. In addition, this component implements two more modules: the "Property matcher" and the "Formaliser".

The "Property matcher" is in charge of finding alignments between the generated predicate, and existing Semantic Web vocabularies. As described in Section 3.5, three main sources are used for retrieving semantic property candidates. For assessing their similarity with the generated predicate a string matching algorithm was implemented, which computes a Levenshtein distance [31] between the IDs of the two predicates. Of course, this component is not intended to be a

contribution to advance the state of the art in ontology matching, its goal is to contribute to a complete implementation of OKE and to provide a possible baseline for comparing results with future improved versions.

Finally, the RDF/OWL writer includes the component "Formaliser". This component implements the formalisation step of the method (cf. Section 3.4). It is in charge of producing the triples summarising the relation expressed in s, and that can be used for annotating the corresponding hyperlink, to generate OWL axioms defining domain and range of the generated predicates, and finally to annotate the produced triples and predicates with scope information.

Legalo for typing Wikipedia pagelinks. A specialised version of Legalo for typing Wikipedia pagelinks (Legalo-Wikipedia)[27] was presented in [38], however it relied on a previous version of the tool. In fact, Legalo-Wikipedia depends on Legalo, hence it evolves with it, and specialises it with two additional features: (i) a sentence extractor specialised for Wikipedia HTML formatting, and (ii) a subject resolver specialised for Wikipedia. A detailed description of this implementation can be found in [38].

Briefly, Legalo-Wikipedia takes in input a DBpedia entity URI, and retrieves all its pagelinks triples from the Pagelinks DBpedia dataset. For each pagelink triple it extracts all Wikipedia snippets containing an hyperlink corresponding to the triple by means of a specialised sentence extractor. Then, the subject resolver selects all and only the snippets that contain a lexicalisation of the Wikipedia page subject, by relying on the DBpedia Lexicalizations Dataset[28].

For example, the wikipage `wp:Ron_Cobb` includes a link to `wp:Sydney` in the sentence:

> "In 1972, Cobb moved to Sydney, Australia, where his work appeared in alternative magazines such as The Digger."

This sentence will be selected and stored as it contains the term "Cobb", which is a lexicalization of `dbpedia:Ron_Cobb`. The same wikipage includes a link to `wp:Los_Angeles_Free_Press` in the sentence:

> "Edited and published by Art Kunkin, the Los Angeles Free Press was one of the first of the underground newspapers of the 1960s, noted for its radical politics."

This sentence will be discarded as it does not include any lexicalisation of `dbpedia:Ron_Cobb`. This procedure is needed for identifying pagelinks that actually convey a semantic factual relation between the Wikipedia page subject and the target of the pagelink. Each snippet is then passed to Legalo as input for generating the Semantic Web property. The version of legalo-wikipedia presented in [38] relied on a previous version of Legalo, which supported less general generative rules and did not perform the relevant relation assessment or the formalisation of the generated property.

5. Results and evaluation

Legalo-Wikipedia has been previously evaluated. For the sake of completeness, these results are summarised in Section 5.2 (for additional details, the reader can refer to [38]). With the help of crowdsourcing an additional, more extensive evaluation of the current implementation of Legalo was performed, which allowed us to better assess its performances and open issues. This section reports this evaluation results in terms of precision, recall, and accuracy.

5.1. Legalo working hypothesis

Legalo is based on two working hypotheses.

Hypothesis 1 (Relevant relation assessment)**.** *Legalo is able to assess if, given a sentence s, a relevant relation exists which holds between two entities, according to the content of s:*

$$\exists\varphi.\varphi_s(e_{subj}, e_{obj})$$

This means that if s contains evidence of a relevant relation between e_{subj} and e_{obj}, then Legalo returns a true value, otherwise it returns false.

Hypothesis 2 (Usable predicate generation)**.** *Legalo is able to generate a usable predicate λ' for a relevant relation φ_s between two entities, expressed in a sentence s: given λ', a label generated by Legalo for φ_s, and λ_i a label generated by a human for φ_s the following holds (cf. Definition 1):*

$$\lambda' \cong \lambda_i, \lambda_i \in \Lambda$$

[27] A demo is available at `http://wit.istc.cnr.it/stlab-tools/legalo/wikipedia`

[28] `http://wiki.dbpedia.org/Datasets/NLP?v=yqj`

which means that the label $\lambda^{'}$ generated by Legalo is equal or very similar to a label λ_i that a human would define in a Linked Data vocabulary for representing φ_s in a particular textual occurrence.

This section reports the evaluation of Legalo based on the validation of Hypothesis 1 and Hypothesis 2.

Evaluation sample. As evaluation data, a corpus $C_{rel-extraction}$ for relation extraction developed at google research[29] was used. There are five datasets available in this corpus, and each dataset is dedicated to a specific relation: place of birth, attending or graduating from an institution, place of death, date of death, degree of education. Each dataset includes a snippet from Wikipedia, a pair (subject, object) of freebase entities, and at least five user judgments that indicate if the snippet contains a sentence providing evidence of a referenced relation (e.g., place of death) between the given pair of entities. It is important to remark that Wikipedia snippets included in the corpus contain more than one sentence, which can be evidence of other relations than the ones for which they were evaluated. Based on this observation, the corpus has been used also for evaluating Legalo on its ability to assess the existence of open-domain relations.

It has to be noticed that Legalo addresses all the capabilities of an OKE system (cf. Section 2), however by using $C_{rel-extraction}$ for its evaluation and considering the homogeneous writing style of Wikipedia authors, additional experiments are needed to properly assess Legalo scalability performance on Web diversity (e.g. blogs, twitter, etc.). In other words, Legalo can be used with any input text, but the different styles of the diverse Web sources could affect its performance. We leave the investigation of possible bias caused by different writing styles to future development.

The evaluation was performed using a subset of $C_{rel-extraction}$. More specifically, three evaluation datasets were derived from $C_{rel-extraction}$ and used for performing different experimental tasks.

- $C_{institution}$: a sample of 130 randomly selected snippets extracted from the file of $C_{rel-extraction}$ dedicated to evidence of relations expressing "attending or graduating from an institution". Legalo was executed on all 130 snippets, including in its input the pair of freebase entities associated with the snippet in $C_{institution}$. For each snippet, Legalo gave an output, either one or more predicates or "no relation evidence" (i.e. false value);
- $C_{education}$: a sample of 130 randomly selected snippets extracted from the file of $C_{rel-extraction}$ dedicated to evidence of relations expressing "obtaining a degree of education". Legalo was executed on all 130 snippets, including in its input the pair of freebase entities associated with the snippet in $C_{education}$. For each snippet Legalo gave always an output, either one or more predicates or "no relation evidence":
- $C_{general}$: a sample of 60 randomly selected snippets extracted from $C_{rel-extraction}$, 15 snippets from each file (excluding "date of death" as Legalo only deals with object properties for the moment). The snippets were broken into single sentences and pre-processed with Tagme [14] in order to enrich them with hyperlinks referring to Wikipedia pages (i.e. DBpedia entities): 186 sentences with at least two recognised DBpedia entities were derived. In total, Legalo produced 867 outputs, of which 262 predicates and 605 "no relation evidence". Notice that the high number of false values is not surprising as in many cases a single sentence may contain a high number of entities, and Legalo had to assess the existence of φ on all possible combinations of pairs.

The resulting triples, predicate formalisations, and scope annotations are accessible via a Virtuoso SPARQL endpoint[30].

There are several works demonstrating that crowdsourcing can be successfully used for building and evaluating semantic resources [17,47,34]. Following these experiences, Legalo was evaluated with the help of crowdsourcing. Five different crowdsourced tasks were defined:

1. assessing if a sentence s provides evidence for the referenced relation (i.e. either "institution" or "education") between two given entities e_{subj} and e_{obj} mentioned in s - based on data from $C_{institution}$ and $C_{education}$, respectively;
2. assessing if a sentence s provides evidence for any relation between two given entities e_{subj} and e_{obj} mentioned in s - based on data from $C_{general}$;

[29] https://code.google.com/p/relation-extraction-corpus/downloads/list

[30] Legalo results can be inspected at http://wit.istc.cnr.it:8894/sparql. The reader can submit a pre-defined default query for retrieving an overview of the dataset.

3. judging if a predicate $\lambda^{'}$ generated by a machine adequately expresses (i.e. it is a good summarisation of) a specific relation (i.e. either "institution" or "education") between two given entities e_{subj} and e_{obj} mentioned in s, according to the content of s - based on data from $C_{institution}$ and $C_{education}$, respectively;
4. judging if a predicate $\lambda^{'}$ generated by a machine adequately expresses (i.e. is a good summarisation of) any relation expressed by the content of s, between two given entities e_{subj} and e_{obj} mentioned in s - based on data from $C_{general}$;
5. creating a phrase λ that summarises the relation expressed by the content of s, between two given entities e_{subj} and e_{obj} mentioned in s - based on data from $C_{general}$.

Task 1 and 2 were used for validating Hypothesis 1. The results of these two tasks were then combined with those from Tasks 3 and 4, for validating Hypothesis 2. Finally, task 5 was used for comparing the similarity between λ values generated by humans and $\lambda^{'}$ values generated by Legalo, for validating Hypothesis 2 from a different perspective.

It is important to remark that Task 1 duplicates the information already available in $C_{rel-extraction}$: this choice was driven by the need for using smaller datasets ($C_{rel-extraction}$ samples) as Legalo evaluation experiments needed to address different evaluation tasks. From an analysis of $C_{rel-extraction}$ it has been noticed that some judgements were incorrect, which can be irrelevant on big numbers while it can bias the results on smaller sets. Hence, the corpus samples were re-evaluated on the evidence task, in order to ensure a high reliability of the judgements. Also, our evaluation focused also on open domain relations, hence addressing a larger number of relations than the one judged originally in the corpus.

The Crowdflower platform[31] was used for conducting the crowdsourcing experiments. All tasks included a set of "gold questions" used for computing a trust score t for each worker. Workers had to first perform their job on 7 test questions, and only those reaching $t > 0.7$ were allowed to continue. The value range of t is $[0, 1]$, the higher the score, the more reliable the worker. Given the strong subjective nature of task 5, only for this task a lower trust score $t > 0.6$ was considered acceptable. Each run of a job for a worker contained 4 questions, and they were free to stop contributing at any time. Each question was performed by at least three workers, in order to allow the computation of inter-rater agreement. More precisely, Table 2 shows how many different workers performed each task, also indicating the hypothesis associated with the task. Besides the initial test questions, in order to keep monitoring workers' reliability, each job contained one test questions. Results from test questions were excluded from the computation of performance measures (i.e., precision, recall, accuracy, agreement).

For tasks 1 and 2, judgements were expressed as "yes" or "no" answers. For tasks 3 and 4, judgments could be assessed on a scale of three values: Agree (corresponding to a value 1 when computing relevance measures), Partly Agree (corresponding to a value 0.5 when computing relevance measures), and Disagree (corresponding to a value 0 when computing relevant measures). Task 5 was completely open. The *confidence* measure is provided by CrowdFlower, it measures the inter-rater agreement between workers weighted by their trust values, hence indicating both agreement and quality of judgements at the same time. It is computed as described in Definition 5[32], and an example is given in Example 5.1:

Definition 5. (Confidence score)
Given a task unit u, a set of possible judgements $\{j_i\}$, with $i = 1, ...n$, a set of trust scores each representing a rater $\{t_k\}$, with $k = 1, ...m$, $t_{sum} = \sum_{k=1}^{m} t_k$ the sum of trust scores of raters giving judgements on u, and $trust(j_i)$ the sum of t_k values of raters that choose judgement j_i, the confidence score $confidence(j_i, u)$ for judgement j_i on the task unit u is computed as follows:

$$confidence(j_i, u) = \frac{trust(j_i)}{t_{sum}}$$

Example 5.1. (Confidence score for evidence judgement)
Table 3 shows the judgements of three raters on the same task unit, where possible judgements are "yes" and "no". $t_{sum} = 0.95 + 0.89 + 0.98 = 2.82$
$confidence(\text{“}yes\text{”}, 582275117) = \frac{0.95+0.98}{2.82} = 0.68$
$confidence(\text{“}no\text{”}, 582275117) = \frac{0.89}{2.82} = 0.31$

When aggregating results for a task unit, the judgement with the higher confidence score is selected. Notice that $confidence(j_i, u) = 1$ when all raters give the same judgement.

[31] http://www.crowdflower.com/

[32] http://success.crowdflower.com/customer/portal/articles/1295977

Hypothesis	Task	#workers
Hypothesis 1	Task 1,2	35
Hypothesis 2	Task 3 (institution)	10
Hypothesis 2	Task 3 (education)	18
Hypothesis 2	Task 4	19
Hypothesis 2	Task 5	12

Table 2: Number of different workers that performed the crowdsourced tasks.

Task unit	Judgement	t
582275117	yes	0.95
582275117	no	0.89
582275117	yes	0.98

Table 3: Example of confidence score computation for a task unit.

Task	Relation	Precision	Recall	F-measure	Accuracy	Confidence
2	Any	**0.83**	0.92	0.87	0.82	0.82
1	Education	**0.95**	0.91	0.93	0.87	0.96
1	Institution	**0.93**	0.90	0.91	0.84	0.94

Table 4: Results of Legalo performance in assessing the evidence of relations between entity pairs in a given sentence s. Performance measures are computed on the judgements collected in Task 1 and 2 based on data from $C_{institution}$, $C_{education}$, and $C_{general}$.

Evaluation of Hypothesis 1 Table 4 shows the results of the evaluation of Hypothesis 1, i.e. Legalo's ability to assess if a sentence s provides evidence of a relation φ_s between two entities (e_{subj}, e_{obj}). Task 1 was designed for evaluating this capability on specific relations, while Task 2 was designed for evaluating this capability on any relation. Each row shows the performance results for a specific run of the task indicating the type of relation tackled and the crowdsourced task.

Legalo's performance is measured by means of standard metrics: precision, recall, f-measure, and accuracy. With the aim of clarifying how to interpret them we briefly report an informal definition of true/false positive, and true/false negative in the context of Tasks 1 to 4. As for Tasks 1-2, given a sentence s, the crowd would say "yes" if a relevant relation exists between a given subject/object pair, and "no" if it does not. Legalo output means "true" (the relation exists) whenever it produces a relation, while it means "false" (the relation does not exist) whenever it does not. Hence, True positive = the number of (true, yes) pairs, False positive = the number of (true, no) pairs, True negative = the number of (false, no) pairs, False negative = the number of (false, yes) pairs.

The results of the crowdsourced tasks demonstrate that the Legalo method has high performance (average F-measure=0.92) on the assessment of $\varphi_s(v_{subj}, v_{obj})$ existence (cf. Hypothesis 1). These results are really satisfactory especially compared with performance results of Legalo-Wikipedia [38], where this aspect was not tackled, and φ_s existence was partly ensured by the nature of input data (cf. see also Section 5.2).

Evaluation of Hypothesis 2 Table 5 shows the results of the evaluation of Hypothesis 2, i.e. Legalo's ability of generating usable predicates for summarising relations between entities, according to the content of a sentence. Task 3 was designed for evaluating this capability on specific properties, while Task 4 was designed for evaluating this capability on any property. Each row shows the performance results indicating the type of relation tackled and the crowdsourced task. The results for "institution" relation and for "any" relation are computed both on the overall set of results, as well as on a subset that ensured a higher confidence rate

(i.e., only results with $confidence(j_i, u) > 0.65$ are included). As far as the evaluation of the "institution" relation is concerned, the subset of results with high confidence is 68% of the whole evaluation dataset, while for "any" relation it is 76%.

For these tasks, positive values (i.e. when Legalo generates a relation, i.e. "true") can be judged by the crowd with "agree", "partly agree" and "disagree". Let A be the number of "agree", PA the number of "partly agree" and D the number of "disagree". As for the negative values, the definition is the same as for Task 1-2, and we reuse their results, as they are on the same datasets. Hence, we compute: True positive = $(A + 0.5 * PA)$, False positive = D, True negative = the number of (false, no) pairs, False negative = the number of (false, yes) pairs.

Finally, Hypothesis 2 was evaluated also by computing a similarity score between human created predicates and Legalo generated ones. Task 5 was performed for collecting at least three labels λ_i for each triple (s, e_{subj}, e_{obj}). As paraphrasing is a highly subjective task, we expected a very low confidence value. Surprisingly, the average confidence value on this task was not that low (0.59). We compared Legalo predicate $\lambda^{'}$ for a triple (s, e_{subj}, e_{obj}) with all λ_i created by the users for that triple. Two different similarity measures were computed: a string similarity score based on Jaccard distance measure[33], and a semantic similarity measure based on the SimLibrary framework [36][34]. The latter is a semantic similarity score that extends string similarity with measures exploiting external semantic resources such as such as WordNet, MeSH or the Gene Ontology. The average Jaccard similarity score between Legalo labels and human ones is 0.63, while the SimLibrary score is 0.80 (the interval value of both scores is $[0, 1]$, the higher the score, the more similar the the two phrases). Before computing the similarity a pre-processing step was performed to the aim of transforming all verbs to their base form and removing all auxiliary verbs from human predicates. The Stanford CoreNLP framework[35] was used to compute the lemma and POS tag of each term in the phrase. This lemmatisation step was necessary in order to ensure a fair comparison of labels based on string similarity as currently Legalo uses only base verb forms.

Also for Hypothesis 2, Legalo shows very satisfactory performance. An impressive result is the high average value of the semantic similarity score (0.80) between user created predicates and Legalo generated ones. This result confirms the hypothesis discussed in [38], saying that the Legalo design strategy was good at producing predicates that are very close to what a human would do when creating a Linked Data vocabulary. In the context of this work, this hypothesis can be extended to the capability to summarise such relations in a way very close to what a generic user would do. This result is very promising from the perspective of evolving Legalo into a summarisation tool, which is one of the envisioned directions of research.

However, by inspecting the different relevance measures, it emerges that while recall is very high on all tasks (0.90 on average), average accuracy is 0.73 and average precision is 0.75. Although these are very satisfactory performances, it is worth identifying the cases that cause the generation of less usable or even bad results. An insight is that lower precision and accuracy are registered especially in the generation of predicates for "institution" (accuracy 0.62, precision 0.65) relations and for "any" relations (accuracy 0.71, precision 0.68) while for "education" relations these measures show significantly higher values (accuracy 0.85, precision 0.92). This turns out as an important lesson learnt. In fact, less satisfactory precision seems due to the fact that many "institution" relations between two entities (X, Z) are described in the form "X received his Y from institution Z" (or similar), i.e. a ternary relation, which in a frame-based representation G corresponds to something like:

```
:receive_1 vn.role:Agent :X ;
:receive_1 vn.role:Theme :Y ;
:receive_1 vn.role:Source :Z .
```

Currently, based on this representation, Legalo would generate a predicate by following the path connecting X to Y, hence without considering the information on Y. The resulting predicate in this case would be "receive from", while a more informative and usable one would clearly be, e.g."receive degree from", assuming that the type of Y is degree. The term degree is an example of a possible type for Y, however whatever is the type of Y, including its type in the predicate would make it much more informative and usable. This case can be easily generalised by exploiting the semantic information about the thematic role that Y plays in participating in the event $receive\_1$. In

[33] Given two strings s_1 and s_2, where c_1 and c_2 are the two character sets of s_1 and s_2, the Jaccard distance J(s_1,s_2) is defined as $J_{sim}(s_1, s_2) = J_{sim}(c_1, c_2) = \frac{|c_1 \cap c_2|}{|c_1 \cup c_2|}$.

[34] http://simlibrary.wordpress.com/

[35] http://nlp.stanford.edu/software/corenlp.shtml

Task	Relation	Precision	Recall	F-measure	Accuracy	Confidence
3	Education	0.92	0.91	0.91	0.85	0.80
3	Institution	0.65	0.91	0.76	0.62	0.59
3 (high confidence only)	Institution	0.74	0.89	0.81	0.68	0.71
4	Any	0.68	0.90	0.78	0.71	0.64
4 (high confidence only)	Any	0.73	0.87	0.80	0.75	0.76

Table 5: Results of Legalo performance in producing a usable label for relations between entity pairs in a given sentence. Performance measures are computed on the judgements collected in Tasks 3 and 4 based on data from $C_{institution}$, $C_{education}$, and $C_{general}$.

fact, a representation pattern can be recognised here: when participating in the event $receive\_1$, X plays an agentive role (as expected from Axiom 3), Y plays a passive role, and Z plays an oblique role. The type of an entity playing a passive role, i.e. Y in this case, is a relevant information as far as the relation between an entity playing an agentive role, and another playing an oblique role in an event, is concerned. This pattern can be generalised to other relations than institution, which explains a similar behaviour of Legalo in the two tasks focusing on assessing usability of predicates for "institution" and "any" relations. Another example that shows this pattern is given by the sentence,

> "Hassan Husseini became an organizer for the Communist Party."

taken from the dataset $C_{general}$. In this case, the representation is the following:

```
:become_1 vn.role:Agent Hassan_Husseini ;
:become_1 vn.role:Patient :Organizer ;
:become_1 :for :Communist_Party .
```

and Legalo would produce the predicate "become for". By applying the new suggested generative rule, the generated predicate would be instead, the more informative and usable "become organizer for". This type of observations leads to the definition of additional generative rules that refine Legalo towards a highly probable improvement on precision and accuracy. New rules are implemented based on the data collected from the evaluation results, hence Legalo demo is constantly evolving.

Evaluating the alignment with existing Semantic Web vocabularies. The matching process performed against LOV, NELL [6], and Watson [10] returned a number of proposed alignments between predicates generated by Legalo and existing properties in Linked Data vocabularies. In order to accept an alignment and include it in the formalisation of a Legalo property p_{new}, a threshold $d_{min} = 0.70$ on the computed similarity score (i.e., normalised difference percentage based Levenshtein distance[36]) was set, i.e. only alignments between properties having $d > 0.70$ were kept for the evaluation. All alignments satisfying this requirements were included in the formalisation of the properties generated during this study[37].

The alignment procedure was executed on 629 Legalo properties p_{new}. For 250 p_{new}, it produced at least one alignment to a Semantic Web property p_{sw} with $d > 0.70$. Three raters independently judged on a scale of three values (Agree, Neutral, Disagree) the resulting alignments based on the available metadata of p_{sw} i.e., comments, labels, domain and range. Table 6 shows the results of the user-based evaluation of the alignments between p_{new} and p_{sw}. The three raters have independently judged the proposed alignment very accurate (Precision 0.84) with a high inter-rater agreement (Kendall's W 0.76). Although it was not possible to compute recall for this evaluation, the low percentage of proposed alignments (only 40%) and the simple method applied suggest that there is considerable room for improvement. This evaluation and the implemented method are to be considered a baseline for future work on this specific task.

5.2. *Results and evaluation of Legalo applied to Wikipedia pagelinks*

A previous study [38] described the evaluation of Legalo-Wikipedia. In this section the results of this evaluation are reported, for the sake of completeness. The main difference between Legalo and its Wikipedia specialised version is that in the latter, the subject of the predicate is always given and there is a high prob-

[36] http://bit.ly/1qd45AQ

[37] All triples, property formalisations, and alignments can be retrieved at http://wit.istc.cnr.it:8894/sparql.

# p_{new} with at least one p_{sw}	Total # of (p_{new}, p_{sw})	Levenshtein threshold	Precision	Kendall's W
250	693	0.7	0.84	0.76

Table 6: Evaluation results on the accuracy of the alignment between p_{new} and p_{sw}.

ability that it is correct based on the design principles that guide Wikipedia page writing. It is worth to remark that the evaluation experiment of Legalo-Wikipedia was performed by Linked Data experts, hence comparing the new results with the previous ones provides insights on the usability of the generated predicates, regardless the expertise of the evaluators.

The evaluation results of Legalo-Wikipedia are published as RDF data and accessible through a SPARQL enpoint[38].

The evaluated sample set consisted of 629 pairs $(s, hyperlink)$, each associated with a FRED graph G. Legalo was executed on this corpus and generated 629 predicates (referred to as p_{new} from now on). The user-based evaluation involved three raters, who are computer science researchers familiar with Linked Data, but not familiar with Legalo. Independently, they have judged the results of Legalo based on two separate tasks, using a Likert scale of five values (Strongly Agree, Agree, Partly Agree, Disagree, Strongly Disagree). When computing performance measures the scale was reduced to three values. Specifically, Strongly Disagree and Agree where associated with a value 1, Partly Agree with 0.5, and Disagree and Strongly Disagree with 0.

The results of the user-based evaluation of p_{new} are reported in Table 7. The three raters have independently judged that the generated predicates p_{new} were very well designed and accurate (F-measure 0.83) in capturing the semantics of their associated pagelinks according to the content of the sentence s, with a high inter-rater agreement (Kendall's W 0.73)[39].

6. Discussion

Dependency on entity linking An aspect that requires improvement is the potential dependency of Legalo performance on the recognition and linking of DBPedia entities in a sentence: if an entity is not in DBpedia, the relation is not generated. Ideally, this is easily solvable by treating any recognised named entity in a sentence as a potential hyperlink, regardless if it has a URI (one can be locally created on the go). A development version of Legalo[40] shows this capability, however besides the need of rigorous experiments for assessing its performance, anecdotical tests show that in some cases this generalisation produces noise in the results. Identifying the causes and handling them is one of our current focus.

Passive form and skolemised entities Identifying recurrent errors helps us identifying new patterns for improving label generation. However, some recurrent mistakes are not easily treatable. One of such cases can be exemplified by the following sentence:

> In March 2008, Evile's track was featured on the Wii, Xbox 360, and PlayStation 3 video game Rock Band as downloadable content.

Currently, Legalo cannot correctly handle this (type of) sentence. There are two main issues motivating this lack: (i) the sentence is expressed in a passive mode, i.e. "was featured on" instead of "features" and the use of the preposition "on" instead of "by" makes the agent "Rock Band" became an oblique role. Hence, there is apparently no agentive role in this sentence, making Axiom 1 (cf. 3) unsatisfied, which causes Legalo to wrongly assess that there is no relevant relation between "Rock Band" and "Evile"; (ii) even if the passive mode was recognised and handled in order to make Axiom 1 satisfied, the target of the passive relation would be "Evile's track", which is a variable i.e. the entity representing it is skolemised. A way to handle this is to name skolemised entities when they show certain characteristics. For example in this case there is a relation between a named entity and the variable and such relation is genitive, hence having a specific recognisable characteristic. However, naming skolemised discourse referent should be done at the level of FRED result as this operation can be useful in

[38] http://isotta.cs.unibo.it:9191/sparql

[39] Kendall's W measures the inter-rater agreement. Values ranges from 0 (complete disagreement) to 1 (complete agreement).

[40] http://wit.istc.cnr.it/kore-dev/legalo

Number of p_{new}	Precision	Recall	F-measure	Kendall's W
629	0.72	0.97	0.83	0.73

Table 7: Evaluation results on the accuracy of p_{new}.

many other application. For example, it can be relevant also for aspect-based sentiment analysis[41].

Open domain and any kind of text sources. The OKE method is meant to support knowledge extraction from text in the open domain. "Open domain" has a twofold interpretation, both valid in this context: (i) any knowledge area: meaning that the approach must be independent from the topics addressed by a text, in other word it should not be tailored to specific languages, vocabularies or terminologies; (ii) any text style: considering that natural language on the Web can have many different writing styles (e.g., a text in a Wikipedia page is certainly *cleaner* than an average blog text, which in turn has a complete different style than twitter writing). The implementation presented in this paper shows very promising results as demonstrated by the performance measured after the execution of a set of crowdsourcing tasks. This evaluation was based on texts extracted from Wikipedia pages, focusing on both specific and general domains, hence showing that the tool works well with any knowledge area. Nevertheless, it remains important to investigate how the change of writing style impacts on the tool performance, in order to assess its behaviour when coping with any text source (going beyond the style of Wikipedia text). This investigation is a main action point in the next evolution of this work. It has to be noticed that Legalo's main tasks are the relation assessment and the label generation, while parsing and role labelling, which are at the base of the frame-based graph representation, are embedded in FRED. In other words, Legalo performance highly depends on the ability of FRED to produce an accurate frame-based representation of the input sentence. This means that intervening for minimising the performance bias due to different writing styles requires to intervene on FRED components (especially parsing and role labelling).

Alignment to existing Semantic Web properties. As for the alignment procedure, there is also space for significant improvement, since this task was addressed by computing a simple Levenshtein distance. More sophisticated alignment methods such as those from the Ontology Alignment Initiative[42] or other approaches for entity linking such as SILK[43] [22] can be investigated for enhancing the alignment results. An interesting result is that our alignment results are good in terms of precision, although all properties that have been matched with a distance score > 0.70 came only from Watson [10] and LOV. We observed that almost all properties retrieved from NELL [6] had an editing distance < 0.70 hence almost none of them were judged appropriate. This reinforces the hypothesis the OKE generative rules simulate very well the results of human property creation, i.e. property names are cognitively well designed. In fact, Watson and LOV are repositories of Semantic Web authored ontologies and vocabularies, while NELL properties result from an artificial concatenation of categories learnt automatically.

As for the alignment recall, it was not possible to compute standard recall metrics because it is impossible to compute False Negative results i.e., all existing Semantic Web properties that would match p_{new} but that we did not retrieve. The relatively high number of missing properties suggests on one hand that a more sophisticated alignment method is needed. On the other hand, if we combine this result with the high value of accuracy of p_{new} and the proposed alignments between p_{new} and p_{sw}, it is reasonable to hypothesise that many cases reveal a lack of intensional coverage in Semantic Web vocabularies, and that OKE can help filling this gap.

Comparison to Open Information Extraction Extracting, discovering, or summarizing relations from text is not an easy task. Natural language is very subtle in providing forms that can express, allude, or entail relations, and syntax offers complex solutions to relate explicitly named entities, anaphoras to mentioned or alluded entities, concepts, and entire phrases, let alone tacit knowledge. Table 8 shows some kinds of (formalisable) relations that can be derived from text.

A full-fledged analysis of those texts is possible to a certain extent, specially if associated with background

[41] http://alt.qcri.org/semeval2015/task12/

[42] http://oaei.ontologymatching.org/

[43] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

sentence	**argument#1**	**binary relation**	**argument#2**
Mr. Miller, 25, entered North Korea seven months ago.	Mr._Miller	enter	North_Korea
He was charged with unruly behavior.	Mr._Miller	charge_with	x:unruly_behavior
North Korean officials suspected he was trying to get inside one of the country's feared prison camps.	y:North_Korean_officials	suspect	try(he, (get_inside (he, z:Korea's_feared_prison_camp)))

Table 8: Sample sentences involving non-trivial relations, expressed in a generic logical form.

knowledge (as FRED does), but the conciseness and directness of hyper-linking based on binary relations is often lost. Hence the importance of tools like Legalo, which are able to reconstruct binary relations from complex machine reading graphs.

It would be natural to compare the results of Legalo to relation extraction systems, but this would require to manipulate their output, which is beyond the scope of this work. Here follows an explanation of the difficulties involved.

A state-of-art tool like Open Information Extraction (OIE, [26]) applies an extractive approach to relation extraction, and solves the problem by extracting segments that can be assimilated to subjects, predicates, and objects of a triplet. As reported in [19], its accuracy was not very high with the version of OIE implemented as the ReVerb tool, but it has sensibly improved recently. However, the segments that are extracted, though useful, are not always intuitively reusable as formal RDF properties or individuals. Table 9 shows one case of a very complex segment #3, i.e. "with a West angry over Russia's actions in Ukraine", which is a phrase to be further analyzed in order to be formalized, and typically leading to multiple triples; and another case of a complex segment #2, i.e. "developed a passion for the native flora of the arid West Darling region identifying", which is not easily transformable into a RDF property.

The research presented here intends to go beyond text segmentation, by using an abstractive approach that selects paths in RDF graphs in order to generate RDF properties. The difference between the two approaches is striking, and leads to results that are difficult to compare. Table 10 shows two of the examples from Table 9 (the third one has no resolvable entity on the object position), but as they are extracted and formalized by Legalo.

For the reasons described above, this work has not attempted a direct comparison in terms of accuracy between OIE and Legalo: it would have needed the transformation and formalization of OIE text segments into individuals and properties, and arbitrary choices on how to formalize complex segments. At the end, it is not a measure of their outputs that is obtained, but a measure of authors' ability to redesign OIE's output. For those interested in attempts to reuse heterogeneous NLP outputs for formal knowledge extraction, see [19].

7. Related Work

The work presented here can be categorised as *formal binary relation discovery and labeling from arbitrary walks in connected fully-labeled multi-digraphs*, which means in practice that it is not just relation extraction (relations are extracted by FRED [39], and Legalo reuses them), but Legalo discovers complex relations that summarise information encoded in several nodes and edges in the graph (RDF graphs are actually connected, fully-labeled multi-digraphs). It considers certain paths along arbitrary directions of edges, aggregating some of the existing labels, and concatenating them in order to provide property names that are typical of Linked Data vocabularies, and finally axiomatizing the properties with domain, range, subproperty, and property chain axioms.

In other words, Legalo tries to answer the following question: what is the relation that links two (possibly distant) entities in a RDF graph?

segment #1	segment #2	segment #3	sentence
Eugene Nickerson	was quarterback of	the football team and captain	*At St. Mark's School in Southborough, Massachusetts, Eugene Nickerson was quarterback of the football team and captain of the hockey team.*
President Vladimir Putin	faced	with a West angry over Russia's actions in Ukraine	*President Vladimir Putin, faced with a West angry over Russia's actions in Ukraine, has been boosting ties to the East.*
Florence May Harding	developed a passion for the native flora of the arid West Darling region identifying	plants	*Early in life Florence May Harding developed a passion for the native flora of the arid West Darling region, collecting and identifying plants.*

Table 9: Some relations extracted bu OIE from sample sentences.

rdf:subject	rdf:property	rdf:object	sentence
`d:Eugene_Nickerson`	`l:quarterbackOf`	`d:American_Football`	*At St. Mark's School in Southborough, Massachusetts, Eugene_Nickerson was quarterback of the football team and captain of the hockey team.*
`d:Vladimir_Putin`	`l:faceWithAngryOverActionLocatedIn`	`d:Ukraine`	*President Vladimir Putin, faced with a West angry over Russia's actions in Ukraine, has been boosting ties to the East.*

Table 10: Two sample extractions by Legalo from the same sentences as in Table 9. For the sake of space we use prefix `d:` instead of `dbpedia:`, and `l:` instead of `legalo:`

There is not much that can be directly comparable in the literature, but work from two related fields can be contrasted with what Legalo does: *relation extraction*, and *automatic summarization*.

The term Open Knowledge Extraction was previously introduced in the context of Artificial Intelligence [11]. This work defines OKE as "conversion of arbitrary input sentences into general world knowledge

represented in a logical form possibly usable for inference", hence perfectly compatible with what defined in this paper. The cited work does not focus on Semantic Web technologies and languages, although it provides further support to our claims and definitions.

The closest works in relation extraction include Open Information Extraction (e.g. [26],[30]), relation extraction exploiting Linked Data [46][24], and question answering on linked data [25].

Relation extraction. The main antecedent to Open Information Extraction is probably the 1999 Open Mind Common Sense project [43], which adopted an ante-litteram crowdsourcing and games-with-a-purpose approach to populate a large informal knowledge base of facts expressed in triplet-based natural language. The crowd was left substantially free to express the subject, predicate, and object of a triplet, but during its evolution, forms started stabilizing, or were learnt by machine learning algorithms. Currently Open Mind is being merged with several other repositories in ConceptNet [21].

Open Information Extraction (aka Machine Reading) as it is currently known in the NLP community performs *bootstrapped* (i.e. started with learning from a small set of seed examples, and then recursively and incrementally applied to a huge corpus, cf. [12]), *open-domain*, and *unsupervised* information extraction. E.g. OIE is based on learning frequent triplet patterns from a huge shallow parsing of the Web, in order to create a huge knowledge base of triplets composed of text chunks.

This idea (on a smaller scale) was explored in [7], with the goal of resolving predicates to, or to enlarge, a biomedical ontology. On the contrary, OIE extracts binary relations by segmenting the texts into triplets. However, there is usually no attempt to resolve the subjects and objects of those triplets, nor to disambiguate or harmonize the predicates used in the triples. Since predicates are not formally represented, they are hardly reusable for e.g. annotating links with RDFa tags. See Section 6 for a comparison between OIE and Legalo, proving the difficulty of even designing a comparison test.

Overall, Open Information Extraction looks like a component for extractive summarization (see below). In [30], named entity resolution is used to resolve the subjects and objects, and there is an attempt to build a taxonomy of predicates, which are encoded as lexico-syntactic patterns rather than typical predicates.

Another important Open Information Extraction project is Never Ending Language Learning (NELL) [6], a learning tool that since 2010 processes the web for building an evolving knowledge base of facts, categories and relations. In this case there is a (shallow) attempt to build a structured ontology of recognised entities and predicates from the facts learnt by NELL. In this work, NELL is used in an attempt to align the semantic relations resulting from Legalo to the NELL ontology.

The main difference between approaches such as OIE and NELL, and Legalo is that the formers focus on extracting mainly direct relations between entities, while Legalo focuses on revealing the semantics of relations between entities that can be: a) directly linked, b) implicitly linked, c) suggested by the presence of links in Web pages, d) indirectly linked, i.e. expressed by longer paths or n-ary relations. Legalo novelty also resides in performing property label generation. From the acquisition perspective, Legalo is not bootstrapped, but it is open-domain and unsupervised.

Relation extraction and question answering targeted at Linked Data are quite different from both Open Information Extraction and Legalo, since they are oriented at formal knowledge, but they are not bootstrapped, open domain and unsupervised. They typically use a finite vocabulary of predicates (e.g. from DBpedia ontology), and use their extensional interpretation in data (e.g. DBpedia) to either link two entities recognized in some text (as in [46][24]), or to find an answer to a question, from which some entities have been recognized (as in [25]). Domain is therefore limited to the coverage of the vocabulary, and distant supervision is provided by the background knowledge (e.g. [1]. A growing repository of relationships extracted with this specific domain, distanly supervised approach is sar-graphs [46].

Automatic summarisation. Automatic summarization deserves a short discussion, since ultimately Legalo's relation discovery can be used as a component for that application task. According to [40], the main goal of a summary is to present the main ideas from one or more documents in less space, typically less than half of one document. Different categorizations of summaries have been proposed: topic-based, indicative, generic, etc., but the most relevant seems to distinguish between "extracts" and "abstracts". Extracts are summaries created by reusing portions of the input text verbatim, while abstracts are created by reformulating or regenerating the extracted content. An

extraction step is needed in any case, but while extracts *compress* the text by squeezing out unimportant material, and *fuse* the reused portions, abstracts typically *model* the text, by accessing external information, applying frames, deep parsing, etc., eventually generating a summary that in principle could contain no word in common with the original text.

Extractive summarization is now in mass usage, e.g. with snippets provided by search engines. It has serious limits, because size and relevance of the extracts can be questionable and not as accurate as a human may be.

Legalo can be considered closer to abstractive summarization, since it can be used to build frame-based abstractive summaries of texts, consisting in binary relation discovery, which can then be filtered for relevance. The current implementation of Legalo is not designed in view of abstractive summarization, therefore it was not evaluated it for that task, but it is appropriate to report at least one relevant example of related work in this area.

Opinosis [18] is the state-of-the-art system for abstractive summarisation. It performs graph-based summarisation, generating concise abstractive summaries of highly redundant opinions. It uses a word graph data structure to represent the text, whereas Legalo uses a semantic graph. As the authors say: "Opinosis is a shallow abstractive summariser as it uses the original text itself to generate summaries. This is unlike a true abstractive summariser that would need a deeper level of natural language understanding". Legalo is indeed based on FRED [39], which provides such deeper level of understanding.

In order to be considered an abstractive summariser, Legalo will need to be complemented with more capabilities to rank discovered relations across an entire or even multiple texts, to associate them in a way that final users can make sense of, and to evaluate summaries appropriately. Results from both abstractive summarisation (e.g. [49][18][23]) and RDF graph summary (e.g. [48][37][5]) can be reused to that purpose.

8. Conclusion and future work

Conclusion. This paper presents a novel approach for Open Knowledge Extraction, and its implementation called *Legalo*, for uncovering the semantics of hyperlinks based on frame-based formal representation of natural language text, and heuristics associated with subgraph patterns. The main novel aspects of the approach are: relevant relation assessment, label generation, Semantic Web property generation and formalisation.

The working hypothesis is that hyperlinks (either created by humans or knowledge extraction tools) provide a pragmatic trace of semantic relations between two entities, and that such semantic relations, their subjects and objects, can be revealed by processing their linguistic traces: the sentences that embed the hyperlinks. Evaluation experiments conducted with the help of a crowdsourcing platform confirm this hypothesis, and show very high performances: the method is able to predict the actual presence of a relation with a high precision (average F-measure 0.92), and generate accurate RDF properties between the hyperlinked entities in single-relation corpora (average F-measure 0.84), the Wikipedia page link corpus (average F-measure 0.84), as well as in the challenging open domain corpus (average F-measure 0.78). The accuracy remains constant across crowdsourced evaluation, and comparison to (crowdsourced) gold standard for the open domain corpus. We also provide alignments to Semantic Web vocabularies with a precision value of 0.84.

A demo of Legalo Web service is available online[13], as well as the prototype dedicated to Wikipedia pagelinks[27], and the binary properties produced in this study can be accessed by means of a sparql endpoint[36].

Ongoing work. Current work concentrates on designing and testing new heuristics, as required by evidence emerging from experiments and tests (cf. e.g. Section 6), on identifying new ways of aligning the relations generated by Legalo to existing ontologies, and on discovering regularities in the relation taxonomies that are increasingly discovered. Additionally, new experiments are under development for assessing Legalo scalability on the diversity and size of the Web.

Future work. The main research line for the future is to apply Legalo to application tasks. An obvious one is a real abstractive summarisation task, both at single-text, and multiple-text level, evaluating the results against state-of-the-art tools. The challenges there include at least: (i) managing multiple (and possibly dynamically evolving) Open Knowledge Extraction graphs, (ii) assessing relevance of discovered relations, and their dependence across a same text, or across multiple texts, and (iii) generating factoid sequences that make sense to a final user of abstractive summaries. Also other applications of Legalo are en-

visioned, including question answering and textual entailment.

References

[1] I. Augenstein, D. Maynard, and F. Ciravegna. Relation extraction from the web using distant supervision. In K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, editors, *Proceedings of Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014*, volume 8876 of *Lecture Notes in Computer Science*, pages 26–41, Linköping, Sweden, 2014. Springer.

[2] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In M. M. Veloso, editor, *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2670–2676, Hyderabad, India, 2007. AAAI Press/International Joint Conferences on Artificial Intelligence.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal of Semantic Web Information Systems*, 5(3):1–22, 2009.

[4] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A crystallization point for the Web of Data. *International Journal of Web Semantics*, 7(3):154–165, 2009.

[5] S. Campinas, T. E. Perry, D. Ceccarelli, R. Delbru, and G. Tummarello. Introducing rdf graph summary with application to assisted sparql formulation. In A. Hameurlain, A. M. Tjoa, and R. Wagner, editors, *Proceedings of the 23rd International Workshop on Database and Expert Systems Applications (DEXA)*, pages 261–266, Vienna, Austria, 2012. IEEE Computer Society.

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In M. Fox and D. Poole, editors, *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI)*, pages 1306–1313, Georgia, USA, 2010. AAAI Press.

[7] M. Ciaramita, A. Gangemi, E. Ratsch, J. Šaric, and I. Rojas. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In L. P. Kaelbling and A. Saffiotti, editors, *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 659–664, Edinburgh, Scotland, 2005. Professional Book Center.

[8] B. Comrie. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press, Chicago, USA, 1989.

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

[10] M. d'Aquin, E. Motta, M. Sabou, S. Angeletou, L. Grindinoc, V. Lopez, and D. Guidi. Towards a new generation of semantic web applications. *IEEE Intelligent Systems*, 23(3):80–83, 2008.

[11] B. V. Durme and L. K. Schubert. Open knowledge extraction using compositional language processing. In R. Basili, J. Bos, and A. Copestake, editors, *Proceedings of the 2008 Conference on Semantics in Text (STEP)*, pages 239–254, Venice, Italy, 2008. The Association for Computational Linguistics.

[12] O. Etzioni, M. Banko, and M. J. Cafarella. Machine reading. In Y. Gil and R. J. Mooney, editors, *Proceedings of the Twenty-first Conference on Artificial Intelligence (AAAI)*, pages 1517–1519, Boston, Massachusetts, 2006. AAAI Press.

[13] J. Euzenat and P. Shvaiko. *Ontology Matching, Second Edition*. Springer, Berlin, Germany, 2013.

[14] P. Ferragina and U. Scaiella. TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, and A. An, editors, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1625–1628, Toronto, Canada, 2010. ACM.

[15] C. J. Fillmore. Frame semantics. In L. S. of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., 1982.

[16] T. Flati, D. Vannella, T. Pasini, and R. Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In Toutanova and Wu [44], pages 945–955.

[17] M. Fossati, C. Giuliano, and S. Tonelli. Outsourcing framenet to the crowd. In P. Fung and M. Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL) Volume 2: Short Papers*, pages 742–747, Sofia, Bulgaria, 2013. The Association for Computational Linguistics.

[18] K. Ganesan, C. Zhai, and J. Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In C. Huang and D. Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 340–348, Beijing, China, 2010. Tsinghua University Press.

[19] A. Gangemi. A Comparison of Knowledge Extraction Tools for the Semantic Web. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *Proceedings of the 10th Extended Semantic Web Conference (ESWC) The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 351–366, Montpellier, France, 2013. Springer.

[20] A. Gangemi and V. Presutti. Ontology Design Patterns. In S. Staab and R. Studer, editors, *Handbook on Ontologies, 2nd Edition*, pages 221–243. Springer Verlag, 2009.

[21] C. Havasi, R. Speer, and J. Alonso. Conceptnet: A lexical resource for common sense knowledge. In N. Nicolov, G. Angelova, and R. Mitkov, editors, *Recent advances in natural language processing V: selected papers from RANLP 2007*, volume 309 of *Current Issues in Linguistic Theory*, pages 269–280. John Benjamins Publishing Company, 2007.

[22] R. Isele and C. Bizer. Active learning of expressive linkage rules using genetic programming. *International Journal of Web Semantics*, 23:2–15, 2013.

[23] H. Ji, B. Favre, W.-P. Lin, D. Gillick, D. Hakkani-Tur, and R. Grishman. Open-domain multi-document summarization via information extraction: Challenges and prospects. In T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, pages 177–201. Springer, Berlin Heidelberg, 2013.

[24] A. Khalili, S. Auer, and A.-C. Ngonga Ngomo. conTEXT – Lightweight Text Analytics using Linked Data. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, and A. Tordai, editors, *The Semantic Web: Trends and Challenges - 11th In-*

ternational Conference, (ESWC). Proceedings, volume 8465 of *Lecture Notes in Computer Science*, pages 628–643, Crete, Greece, 2014. Springer.

[25] V. Lopez, A. Nikolov, M. Sabou, V. S. Uren, E. Motta, and M. d'Aquin. Scaling up question-answering to linked data. In P. Cimiano and H. S. Pinto, editors, *Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses (EKAW)*, pages 193–210, Lisbon, Portugal, 2010. Springer.

[26] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni. Open language learning for information extraction. In Tsujii et al. [45], pages 523–534.

[27] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In K.-Y. Su, editor, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, pages 1003–1011, Suntec, Singapore, 2009. Association for Computational Linguistics.

[28] A. Moro and R. Navigli. Integrating syntactic and semantic analysis into the open information extraction paradigm. In F. Rossi, editor, *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2148–2154, Beijing, China, 2013. AAAI Press/IJCAI.

[29] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244, 2014.

[30] N. Nakashole, G. Weikum, and F. Suchanek. Patty: A taxonomy of relational patterns with semantic types. In Tsujii et al. [45], pages 1135–1145.

[31] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[32] A. G. Nuzzolese, A. Gangemi, and V. Presutti. Gathering Lexical Linked Data and Knowledge Patterns from FrameNet. In M. A. Musen and Ó. Corcho, editors, *Proceedings of the sixth international conference on Knowledge Capture (K-CAP)*, pages 41–48, Banff, AB, Canada, 2011. ACM.

[33] A. G. Nuzzolese, A. Gangemi, V. Presutti, and P. Ciancarini. Encyclopedic Knowledge Patterns from Wikipedia Links. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, and E. Blomqvist, editors, *Proceedings fo the 10th International Semantic Web Conference (ISWC), Part I*, volume 7031 of *Lecture Notes in Computer Science*, pages 520–536, Bonn, Germany, 2011. Springer.

[34] J. Oosterman, A. Nottamkandath, C. Dijkshoorn, A. Bozzon, G. Houben, and L. Aroyo. Crowdsourcing knowledge-intensive tasks in cultural heritage. In F. Menczer, J. Hendler, W. H. Dutton, M. Strohmaier, C. Cattuto, and E. T. Meyer, editors, *ACM Web Science Conference (WebSci)*, pages 267–268, IN, USA, 2014. ACM.

[35] S. Peroni, A. Gangemi, and F. Vitali. Dealing with markup semantics. In C. Ghidini, A. N. Ngomo, S. N. Lindstaedt, and T. Pellegrini, editors, *Proceedings the 7th International Conference on Semantic Systems (I-SEMANTICS)*, ACM International Conference Proceeding Series, pages 111–118, Graz, Austria, 2011. ACM.

[36] G. Pirrò and J. Euzenat. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *Proceedings of the 9th International Semantic Web Conference (ISWC) Part I*, pages 615–630, 2010.

[37] V. Presutti, L. Aroyo, A. Adamou, A. Gangemi, and G. Schreiber. Extracting core knowledge from linked data. In O. Hartig, A. Harth, and J. Sequeda, editors, *Proceedings of the Second International Workshop on Consuming Linked Data (COLD)*, volume 782 of *CEUR Workshop Proceedings*, Bonn, Germany, 2011. CEUR-WS.org.

[38] V. Presutti, S. Consoli, A. G. Nuzzolese, D. R. Recupero, A. Gangemi, I. Bannour, and H. Zargayouna. Uncovering the semantics of wikipedia pagelinks. In K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, editors, *Knowledge Engineering and Knowledge Management - Proceedings of the 19th International Conference (EKAW)*, volume 8876 of *Lecture Notes in Computer Science*, pages 413–428, Linköping, Sweden, 2014. Springer.

[39] V. Presutti, F. Draicchio, and A. Gangemi. Knowledge extraction based on Discourse Representation Theory and Linguistic Frames. In A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Aquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, editors, *Knowledge Engineering and Knowledge Management - Proceedings of the 18th International Conference (EKAW)*, volume 7603 of *Lecture Notes in Computer Science*, pages 114–129, Galway City, Ireland, 2012. Springer.

[40] D. R. Radev, E. Hovy, and K. McKeown. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408, 2002.

[41] G. Rizzo, R. Troncy, S. Hellmann, and M. Bruemmer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *Proceedings of the 5th Workshop on Linked Data on the Web (LDOW) co-located with the International World Wide Web Conference (WWW)*, volume 937 of *CEUR Workshop Proceedings*, Lyon, France, 2012. CEUR-WS.org.

[42] K. K. Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2006.

[43] P. Singh. The public acquisition of commonsense knowledge. In J. Karlgren, editor, *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, Palo Alto, CA, USA, 2002. AAAI Press.

[44] K. Toutanova and H. Wu, editors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, (ACL) Volume 1: Long Papers*, Baltimore, Maryland, 2014. The Association for Computer Linguistics.

[45] J. Tsujii, J. Henderson, and M. Pasca, editors. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Jeju Island, Korea, 2012. The Association of Computational Linguistics.

[46] H. Uszkoreit and F. Xu. From strings to things sar-graphs: A new type of resource for connecting knowledge and language. In S. Hellmann, A. Filipowska, C. Barrière, P. N. Mendes, and D. Kontokostas, editors, *Proceedings of the NLP & DBpedia workshop co-located with the 12th International Semantic Web Conference (ISWC 2013)*, volume 1064 of *CEUR Workshop Proceedings*, Sydney, Australia, 2013. CEUR-WS.org.

[47] D. Vannella, D. Jurgens, D. Scarfini, D. Toscani, and R. Navigli. Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. In Toutanova and Wu [44], pages 1294–1304.

[48] X. Zhang, G. Cheng, and Y. Qu. Ontology summarization

based on rdf sentence graph. In C. Williamson and M. E. Zurko, editors, *Proceedings of the 16th International conference on World Wide Web*, pages 707–716, Banff, Alberta, 2007. ACM.

[49] L. Zhou, C.-Y. Lin, D. S. Munteanu, and E. Hovy. ParaEval: Using Paraphrases to Evaluate Summaries Automatically. In R. C. Moore, J. A. Bilmes, J. Chu-Carroll, and M. Sanderson, editors, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 447–454, New York, New York, 2006. The Association for Computational Linguistics.