Uncovering the semantics of Wikipedia pagelinks

Valentina Presutti[1], Sergio Consoli[1], Andrea Giovanni Nuzzolese[1]
Diego Reforgiato Recupero[1] Aldo Gangemi[1,2]
Ines Bannour[2], and Ha ifa Zargayouna[2]

[1] STLab-ISTC Consiglio Nazionale delle Ricerche, Rome, Italy.
[2] Universit Paris 13 - Sorbonne Paris Cité - CNRS, France.

Abstract. Wikipedia pagelinks, i.e. links between Wikipages, carry an intended semantics: they indicate the existence of a factual relation between the DBpedia entity referenced to by the source Wikipage, and the DBpedia entity referenced to by the target Wikipage of the link. These relations are represented in DBpedia as triple occurrences of a generic "wikiPageWikilinks" property. We designed and implemented a novel method for uncovering the intended semantics of pagelinks, and represent them as semantic relations. In this paper, we experiment our method on a subset of Wikipedia showing its potential impact on DBpedia enrichment.

1 Introduction

Wikipedia proved to be an extremely valuable resource for the Semantic Web. In fact, DBpedia (its RDF version) is the most important hub of Linked Open Data [1] and one of the resources most used as training/background knowledge for Semantic Web methods.

However, DBpedia only includes structured knowledge from Wikipedia infoboxes, and categories. There is huge amount of knowledge in Wikipedia that is only expressed as natural language content, which is the case also for the Web in general, and that is worth to annotate with structured data, e.g. by means of RDFa, and transform to Linked Data for fostering the development of the Semantic Web.

Current knowledge extraction (KE) systems address very well the task of linking pieces of text to Semantic Web entities (e.g. `owl:sameAs`) by means of named entity recognition (NER) methods, e.g. NERD[3] [7], FOX[4], conTEXT[5] [9], Dbpedia Spotlight[6], Stanbol[7]. Some of them (e.g. NERD) also perform sense tagging, adding knowledge about entity types (`rdf:type`). However, Wikipedia (and DBpedia) can be enriched also with other semantic relations than `owl:sameAs` and `rdf:type`, i.e. factual relations between entities.

[3] `http://nerd.eurecom.fr`
[4] `http://aksw.org/Projects/FOX.html`
[5] `http://context.aksw.org/app/`
[6] `http://dbpedia-spotlight.github.com/demo`
[7] `http://stanbol.apache.org`

A *pragmatic trace* of a semantic relation between two entities in Wikipedia is the presence of links. In fact, when we include a link in a Wikipage, we usually have a semantic relation in mind between the entity referred to by the page, i.e. subject, and the entity referred to by the target page, i.e. object. For example, a link to "French Open" in the Wikipedia page of "Paris" suggests a semantic relation between these two entities.

This hypothesis is also supported by a previous study [12], in which we have extracted encyclopedic knowledge patterns for DBpedia types, based on links between Wikipedia pages. In [12], a user-based study showed that pagelinks between Wikipedia pages determine relevant descriptive contexts for DBpedia entities at the type level, which suggests that pagelinks mirror relevant semantic relations between entities.

Revealing the semantics of pagelinks has a high potential impact on the amount of Wikipedia knowledge that can be published in machine readable form, keeping the binding with its corresponding natural language expressions. Notably, Wikipedia includes ∼136.6M[8] links between wikipages, i.e. pagelinks. These pagelinks are represented as triples of a generic RDF property "`dbpo:wikiPageWikiLink`"[9]. Our aim is to type pagelinks with the semantic relations that they implicitly convey, by making them explicit. How to do this automatically?

We can use the text surrounding pagelinks. This, in addition to the pragmatic trace, i.e. a link, provides us with a *linguistic trace* of such semantic relations. In fact, the text within which we include a link usually expresses directly or indirectly its intended semantics. For example, the sentence:

"Paris hosts the annual French Open Grand Slam tennis tournament."

explains the semantics of the link between the entity "Paris" and "French Open". Currently, DBpedia represents such link as follows:

`dbpedia:Paris`[10] `dbpo:wikipagewikilink dbpedia:French_Open`

Our aim is to automatically annotate these links with semantic relations based on the sentences expressing their meaning. As for this example, we want to produce something like:

`dbpedia:Paris myont:hostsTournament dbpedia:French_Open`

and when possible to align the extracted property `myont:hostsTournament` to an existing semantic Web property, e.g.:

`myont:hostsTournament rdfs:subPropertyOf gnd:organizerOrHost`[11]

In this paper, we describe a novel approach, named *Legalo*, for automatically typing pagelinks, and we evaluate its performance on a subset of Wikipedia. Legalo is based on a pipeline of components, which has at its core a machine reader, i.e. FRED[12] [14]. FRED transforms natural language text to a formal

[8] M stands for millions.

[9] Prefix `dbpo:` stands for http://dbpedia.org/ontology/

[12] `http://wit.istc.cnr.it/stlab-tools/fred/`

RDF/OWL graph representation. Legalo implements a set of graph pattern-based rules for extracting, from FRED graphs, Semantic Web binary relations that capture the semantics of specific links. This approach can be generalized and applied to any web page and its links[13].

The contribution of this paper can be summarized as follows:

- a novel method for automatic typing Wikipedia pagelinks, which performs property label generation, and graph-based relation extraction;
- an online tool named *Legalo*, which implements and demonstrates such method;
- a user-based evaluation of Legalo on a subset of Wikipedia.

The paper is structured as follows. In Section 2 we discuss related work. Section 3 illustrates the data sources that we have used in our study, while Section 4 describes our approach for automatic typing pagelinks. Section 5 presents our results and the evaluation of our approach. Finally, Section 6 discusses results and future developments.

2 Related work

The closer domain to our research is relation extraction. There are a number of valuable contributions in this field. Notably, Open Information Extraction (OIE)[14] is a project whose aim is to parse the Web and build a knowledge base of triplets extracted from recognized facts. One of the main tools developed in this project is Ollie [10]. Typical results of relation extraction tools such as OIE are not represented as RDF properties, making them hardly reusable for e.g. annotating links with RDFa tags. A possible reuse of OIE in the context of Legalo is to exploit its knowledge base as a lexical resource providing labels for designing more accurate naming strategies for semantic relations, at least for direct connections among entities.

NELL[15] [3] is a learning tool that since 2010 processes the web for building an evolving knowledge base of facts, categories and relations. We used NELL in our work in an attempt to align the semantic relations resulting from Legalo to NELL ontology.

The main difference between approaches such as OIE and NELL (and in general relation extraction tools), and Legalo is that the formers focus on extracting mainly direct relations between entities, while Legalo focuses on unrevealing the semantics of relations between entities that are suggested by the presence of links in Web pages, and that can be also indirect i.e., expressed by longer paths or n-ary relations. Legalo novelty also resides in performing property label generation, and graph-based relation extraction.

FRED[16] [14] is a tool that transforms natural language text in RDF/OWL graphs with an event-based representation approach. Both NELL and FRED are

[13] A possible journal version of this paper would include a generalized version of Legalo, some improvements in its main algorithm, and additional evaluation.

[14] `http://openie.cs.washington.edu/`

[15] `http://rtw.ml.cmu.edu/rtw/`

[16] `http://wit.istc.cnr.it/stlab-tools/fred/`

machine readers although they use different approaches. FRED is based on Discourse Representation Theory (DRT) and Frames [2], and builds its knowledge bases by formalizing natural language text. FRED graphs captures very well the semantics of natural language sentences by performing a deep text analysis. As discussed in Section 4, Legalo builds on top of FRED for synthesizing binary relations representing the semantics of pagelinks, hence responding to a requirements of better cognitive ergonomics than n-ary relations, which are the typical representation patterns in FRED graphs.

Legalo includes a matching approach for identifying existing Semantic Web properties that can be aligned to its produced binary relations, for disambiguation purposes. In this research work, we focused more on the property name design strategies for producing binary relations than on the alignment procedure, although our results are reasonably satisfying also in this respect (cf. Section 5). Most ontology alignment methods [5] (cf. see the Ontology Alignment Evaluation Initiative[17]) work on comparing and aligning collection of ontologies (typically addressing the same knowledge domain). Our task instead is closer to string matching as our input values are two properties accompanied with, in the best case, a `rdfs:comment` describing their intended semantics, and sometimes having defined domain and range. In future work, we intend to investigate how to exploit existing ontology matching methods as well as entity linking tools such as SILK [8] in attempt to improve alignment results.

3 Data sources

In the context of this work we have used a number of data sources.

Wikipedia and Wikipedia Pagelinks. Wikipedia is a collaboratively built encyclopedia on the Web. Currently, English Wikipedia contains ∼4.5M articles[18]. Each Wikipedia page refers to one entity: these entities are represented in DBpedia, the RDF version of Wikipedia. The Wikipedia Pagelinks dataset[19] represents internal links between DBpedia instances as they occur in their corresponding Wikipedia pages. This dataset counts ∼136.6M `dbpo:wikiPageWikiLink` triples (as of version 3.9). We use a subset of Wikipedia pages and their pagelinks as testing sample for evaluating our approach.

Given a sentence including a link, our method creates a RDF property synthesizing the link's semantics. When possible, it aligns this property to existing Semantic Web properties retrieved from three different sources:

– **Watson**[20] [4] is a service that provides access to Semantic Web knowledge, in particular ontologies;

[17] `http://oaei.ontologymatching.org/`

[18] Source: `http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia`, July 2014

[19] `http://wiki.dbpedia.org/Downloads39`

[20] `http://watson.kmi.open.ac.uk/WatsonWUI/`

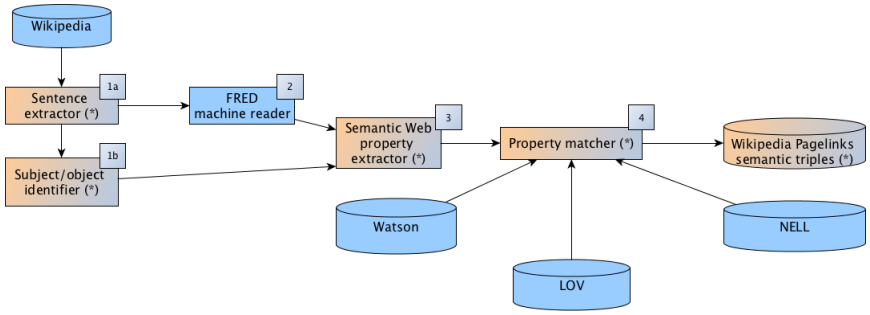

Fig. 1: Pipeline implemented by Legalo for automatic typing Wikipedia pagelinks based on their linguistic trace, i.e. natural language sentence including links. Numbers indicate the order of execution of a component in the pipeline. The output of a component i is passed as input to the next $i+1$ component. (*) denotes tools developed in this work, which are part of our contribution.

- **Linked Open Vocabularies (LOV)**[21] is an aggregator of Linked Open vocabularies (including DBpedia), and provides services for accessing their data;
- **Never-Ending Language Learning (NELL)**[22] [3] 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[23].

4 Legalo: a method and tool for typing hyperlinks

Legalo[24] is based on a pipeline of components and data sources, executed in the sequence illustrated in Figure 1. In [6], we experimented the same pipeline for automatic typing DBpedia entities, and applied it on the whole Wikipedia producing an ontology of Wikipedia entity types, derived from their natural language definitions [13]. Such result supports our hypothesis that the same approach can show promising results for automatic typing of pagelinks.

1a. Sentence extractor. Given a DBpedia entity $subj_e$ (and its corresponding Wikipedia page $page(subj_e)$), this component collects its pagelinks triples (from the Wikipedia Pagelinks dataset, cf. Section 3). For each triple:

$$subj_e \texttt{ dbpo:wikiPageWikiLink } obj_e$$

it extracts the natural language sentences in $page(subj_e)$ including links to $page(obj_e)$, by performing the following actions.

[21] http://lov.okfn.org/dataset/lov/

[22] http://rtw.ml.cmu.edu/rtw/

[23] http://nell-ld.telecom-st-etienne.fr/

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

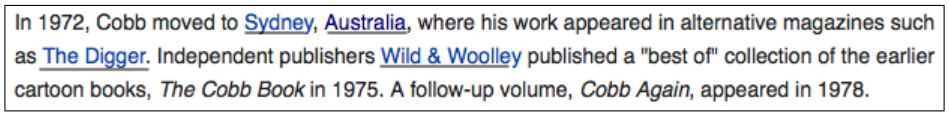

Fig. 2: A fragment of the wikipage `wp:Ron_Cobb` including two sentences with links.

First, it performs a cleaning procedure on $page(subj_e)$: removing special characters, infoboxes, tables, lists, external references, pictures, and captions.

Then, for each pagelink, it extracts the text starting after a dot, and ending with a dot, which includes a link to $page(obj_e)$. Sentences are stored in a RDF graph by associating them to the pair $(subj_e, obj_e)$.

All red links and disambiguation pages are discarded; redirect pages are handled by substituting their URIs with the redirected objects.

For example, given the fragment of the page `wp:Ron_Cobb`[25] depicted in Figure 2[26] the sentence extractor will store the data shown in Table 1.

link ID	$subj_e$	obj_e	**sentence**
L1	`dbpedia:Ron_Cobb`	`dbpedia:Sydney`	In 1972, Cobb moved to Sydney, Australia, where his work appeared in alternative magazines such as The Digger.
L2	`dbpedia:Ron_Cobb`	`dbpedia:Australia`	In 1972, Cobb moved to Sydney, Australia, where his work appeared in alternative magazines such as The Digger.
L3	`dbpedia:Ron_Cobb`	`dbpedia:The_Digger_(alternative_magazine)`	In 1972, Cobb moved to Sydney, Australia, where his work appeared in alternative magazines such as The Digger.

Table 1: Data extracted by the sentence extractor from the wikipage fragment shown in Figure 2.

1b. Subject/object identifier. This component has the role of identifying subject and object of a semantic relation suggested by a pagelink, and their lexicalizations in the associated sentence.

In Wikipedia (and Wikis in general) each $page(subj_e)$ is a representative for one entity, i.e. $subj_e$, hence it is reasonable to assume that all its links suggest semantic relations having $subj_e$ as subject and the links' targets as objects. In this case, we know a priori the DBpedia entities playing the roles of subject and object, respectively: they are given by the sentence extractor as shown above.

Although this assumption is reasonable, there are pagelink sentences that refer to other entities than $subj_e$, but still they are associated with pagelinks of $subj_e$. We identify and keep only those sentences (and their associated pagelinks) that include an explicit lexical reference $lex(subj)$ to $subj_e$. To this aim we use the DBpedia Lexicalizations Dataset[27]. 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."

[25] Prefix wp: stands for http://en.wikipedia.org/wiki/

[26] Links are shown as underlined text.

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

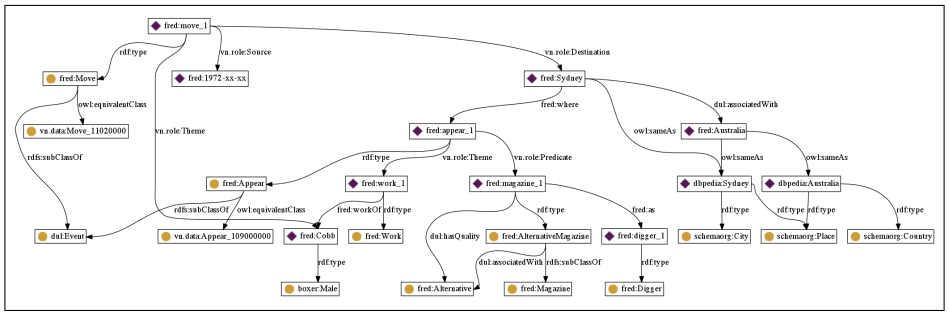

Fig. 3: FRED graph for the sentence: *"In 1972, Cobb moved to Sydney, Australia, where his work appeared in alternative magazines such as The Digger."*

This sentence will be kept and stored in our dataset 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 lexicalization of `dbpedia:Ron_Cobb`. The RDF dataset shown in Table 1 will be updated by this component with the following data:

link ID	*lex*(*subj*)	*lex*(*obj*)
L1	Cobb	Sydney
L2	Cobb	Australia
L3	Cobb	The Digger
L4	Cobb	Wild & Woolley

Table 2: Data extracted by the sentence extractor from the wikipage fragment shown in Figure 2.

2. Machine reading (FRED). Each stored sentence is parsed by FRED[28] [14], which produces its RDF/OWL graph representation. The resulting graph is stored as a named graph and associated with its corresponding pair (pagelink, sentence). FRED implements an event-based representation of natural language text. Hence, relations between entities are mainly represented as n-ary relations. For example, the FRED graph for the sentence:

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

is depicted in Figure 3. Let us consider the pagelink connecting "Cobb" and "Sydney". The resulting graph correctly represents an event occurrence `fred:move_1`[29]

[28] FRED is available online at `http://wit.istc.cnr.it/stlab-tools/fred`

[29] Prefix fred: stands for http://www.ontologydesignpatterns.org/fred/domain.owl - this is the namespace of ontology entities learnt by FRED and can be customized by users.

of type `fred:Move` (i.e. n-ary relation), having `fred:Cobb` and `fred:Sydney` as arguments. If we consider the pagelink connecting "Cobb" and "The Digger" we can find a semantic connection between them within the FRED graph involving an event occurrence of type `fred:Appear` having two arguments (representing Cobb's work and alternative magazines) that are respectively connected to `fred:Cobb` and `fred:Digger`.

On one hand, this event-based representation of a sentence is excellent if we give priority to completeness and richness aspects. On the other hand, from a cognitive ergonomics perspective, it is desirable (when possible) to represent such connections as binary relations. Binary relations provide a semantic "cognitive shortcut" addressing many Linked Data applications; additionally, they can be used for annotating pagelinks by means of RDFa. The aim of Legalo is to synthesize such connections as binary relations.

3. Semantic Web property extractor. This component is in charge of extracting from FRED graphs, the subgraphs connecting $subj_e$ and obj_e, and generating a binary relation expressing the semantics of their connection. One of the main novel aspects of Legalo resides in this component, which performs property label generation and graph-based relation extraction.

Elements' IDs in FRED graphs reflect the terms used in the text, modulo a normalization performed by means of a stemming process. FRED graphs include a set of metadata triples[30] that associate each graph entity to its corresponding text spans.

For example, the term "Australia", starting from the text span "31" and ending at text span "40" in the sentence *"In 1972, Cobb moved to Sydney, Australia, where his work appeared in alternative magazines such as The Digger."* denotes the entity `fred:Australia` in the graph depicted in Figure 3. The following triples provide this information[31].

```
fred:offset_31_40_Australia
a  pos:PointerRange ;
       rdfs:label "Australia"^^xmls:string ;
      semion:denotes fred:Australia ;
       pos:begins "31"^^xmls:nonNegativeInteger ;
       pos:ends "40"^^xmls:nonNegativeInteger ;
```

The Semantic Web property extractor component, for each pair (pagelink, sentence) takes as input the lexicalizations of subject and object, and the associated named graph. Referring to the above example it will receive the data shown in Table 3.

The component performs the following actions: (i) it identifies the nodes denoted by $lex(subj)$ and $lex(obj)$ in the FRED graph, using the text span metadata; (ii) it extracts all subgraphs connecting the two identified nodes; (iii) it produces a Semantic Web (binary) property.

---

[30] These triples are not returned in the graph-view result of FRED at `http://wit.istc.cnr.it/stlab-tools/fred/`, they are returned with all other serialization output options.

[31] Prefix `pos:` stands for `http://www.essepuntato.it/2008/12/earmark#`, `semion:` stands for `http://ontologydesignpatterns.org/cp/owl/semiotics.owl#`, and `xmls:` stands for `http://www.w3.org/2001/XMLSchema#`

link ID	sentence	*lex*(*subj*)	*lex*(*obj*)	**NG ID**
L2	In 1972, Cobb moved to Sydney, Australia, where his work appeared in alternative magazines such as The Digger.	Cobb	Australia	NG1201[32]

Table 3: Example of input data to the Semantic Web property extractor for the pagelink (`dbpedia:Ron_Cobb, dbpedia:Australia`) associated with the linguistic trace "In 1972, Cobb moved to Sydney, Australia, where his work appeared in alternative magazines such as The Digger."

To design the strategies implemented by this component for step (iii), we used a combined top-down and bottom-up approach.

The top down approach consisted in concatenating all labels of graph entities in the extracted subgraphs belonging to the `fred:` namespace, including the more general types of intermediate nodes. With this strategy, each link was associated to a number of properties resulting from different possible paths, each having different length.

The bottom-up approach consisted in evaluating a sample of ∼200 cases, which showed that the minimum path lengths produced the best names for properties. Also, by means of this empirical observation we noticed that when subjects and objects are connected through an event node, the best strategy is to exclude all labels of arcs and types that connect the subject to the event node[33].

Additionally, we associate some of the thematic roles of events' arguments to specific labels. For example, in the above example, the entity `fred:Australia` plays the role `vn.role:Destination`[34], which is associated to the term "to". The association between thematic roles and labels is shown in Table 4.

Thematic role	**label**	**Thematic role**	**label**
vnrole:Actor1	with	vnrole:Actor2	with
vnrole:Agent	by	vnrole:Beneficiary	for
vnrole:Destination	to	vnrole:Instrument	with
vnrole:Topic	about		

Table 4: Labels associated with thematic roles used for producing binary relation names.

As a result, referring to the example above, the produced triple is:

```
dbpedia:Ron_Cobb fred:moveTo dbepdia:Australia
```

4. Property matcher. The aim of this component is to align the produced properties to existing properties defined in Semantic Web ontologies and vocabularies. As described in Section 3, we used three main sources for retrieving semantic property candidates. For assessing their similarity with the property

[33] A detailed description of the algorithm can be temporarily found at http://stlab.istc.cnr.it/stlab/Legalo, which would be included in a possible extended journal version.

[34] Prefix `vn.role:` stands for `http://www.ontologydesignpatterns.org/ont/vn/abox/role/`

produced by the Semantic Web property extractor component, we implemented a string matching algorithm, which computes a Levenshtein distance metrics [11] between the two property IDs.

5 Result and evaluation

In this section, we report the results of our work and evaluate them in terms of precision and recall.

Legalo[35] implements the approach described in Section 4. The goal of this study is to evaluate its potential and identify weaknesses to guide its improvement, towards its application on large scale data, i.e. the whole Wikipedia.

To this aim, we have performed an experiment on a subset of Wikipedia pages and their corresponding pagelinks, taken from the Wikipedia Pagelinks dataset[36]. The experiment results are published as RDF data and accessible through a SPARQL enpoint[37], and have been evaluated by means of a user study described in the remaining of this section.

Building the evaluation sample. To build our experimental setting we have randomly selected 1,000 distinct DBpedia entities, and their corresponding Wikipedia pages, having at least 10 Wikipedia pagelinks[38]. The random selection has been run for each top-level type of the DBpedia ontology (version 3.9) by considering the frequency of instances *per* each of these types on the overall DBpedia.

We have extracted from the collected Wikipedia pages all sentences embedding the selected Wikipedia pagelinks and stored them together with their corresponding pairs ($subj_e$, obj_e) in a RDF triplestore based on Virtuoso[39].

During this process a number of pagelinks have been discarded because we could not find a sentence embedding the link. For example, this happens when a link is found in a table, list, or infobox. From the resulting data, we only kept entries including a sentence with an explicit lexicalization of $subj_e$.

The result of this process led us to collect 1,192 pairs (pagelink, sentence), corresponding to 442 distinct subject entities, i.e. $subj_e$. It has to be noted that discarded sentences sometimes express a relation between $subj_e$ and obj_e. However, they include an indirect reference to $subj_e$, for example by means of anaphoras.

Of these 1,192 sentences we kept only those treatable with FRED, i.e. those for which FRED returned a non-empty graph[40], and having at least a sub-graph connecting $subj_e$ and obj_e either through a direct path of any length, or through an event occurrence.

The final sample set consists of 629 pairs (pagelink, sentence) each associated with a FRED graph.

---

[35] http://wit.istc.cnr.it/stlab-tools/legalo/wikipedia

[36] http://wiki.dbpedia.org/Downloads39

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

[38] We empirically noticed in [12] that this guarantees enough text in the page.

[39] http://virtuoso.openlinksw.com/

[40] Empty graph output from FRED is mainly caused by the presence of not-well formatted characters.

Binary relations for pagelinks. We ran Legalo on the corpus described above, and this process produced 629 new binary relations (referred to as p_{new} from now on), which are retrievable through a SPARQL endpoint[41].

The matching process performed against LOV[42], NELL[43] [3], and Watson[44] [4] returned a number of proposed alignment properties (referred to as p_{sw} from now on) for 250 distinct p_{new}, with a threshold on the editing distance value (Levenshtein distance), normalized as *difference percentage*[45], set to 0.7.

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.

User-based evaluation of p_{new}**.** The first task consisted in judging a new produced property p_{new} based on a pagelink and its associated sentences. For each p_{new} raters have been provided with:

- $lex(subj)$
- $lex(obj)$
- sentence
- p_{new}

$lex(subj)$	$lex(obj)$	**sentence**	p_{new}
Cobb	Spain	While Cobb was in Spain working on Conan the Barbarian, Spielberg supervised the rewrite into the more personal E.T. and ended up directing it himself.	fred:locatedIn

Table 5: An example of evaluation entry for p_{new}.

Table 5 shows an example of evaluation entry for p_{new}. The following evaluation guidelines have been provided to the raters:

> Express your judgment on the following statement, by assigning one of the values shown in Table 6 to each row.
> *"Property p_{new} captures the essential semantics of the relation between subject and object as it is expressed by the sentence."*

The results of the user-based evaluation of p_{new} are reported in Table 7. The three raters have independently judged the produced properties p_{new} very well designed and accurate (F-measure 0.83) in capturing the semantics of their associated pagelinks and linguistic trace, with a high inter-rater agreement (Kendall's

[41] http://isotta.cs.unibo.it:9191/sparql
[42] http://lov.okfn.org/dataset/lov/
[43] http://rtw.ml.cmu.edu/rtw/
[44] http://watson.kmi.open.ac.uk/WatsonWUI/
[45] bit.ly/1qd45AQ

Likert value	Judgement criteria
Strongly agree	The property is very well defined.
Agree	The property captures the relation semantics but the name can be improved.
Neutral	The property captures only partly the relation semantics but can be used.
Disagree	The property does not really capture the semantics of the relation.
Strongly Disagree	The property expresses something completely different from the relation semantics.

Table 6: Likert scale values and associated criteria for evaluating p_{new}

W 0.73). This result suggests that Legalo approach for automatically designing Semantic Web properties satisfies the cognitive-based requirements of this task, i.e. to generate a property name acceptable to a human for expressing a given intensional semantics.

Number of p_{new}	**Precision**	**Recall**	**F-measure**	**Kendall's W**[46]
629	0.72	0.97	0.83	0.73

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

User-based evaluation of alignments (p_{new}, p_{sw}). The second evaluation task consisted in judging the proposed alignments to existing Semantic Web properties p_{sw} based on the information provided by their metadata i.e., comments and labels, and their domain and range (when available).

For each pair (p_{new}, p_{sw}) raters have been provided with the data of the previous evaluation entries extended with:

- p_{sw}
- `rdfs:comment` of p_{sw}
- `rdfs:label` of p_{sw}
- `rdfs:domain` of p_{sw}
- `rdfs:range` of p_{sw}

p_{new}	p_{sw}	`rdfs:comment`	`rdfs:label`	`rdfs:domain`	`rdfs:range`
`fred:locatedIn`	`geo:locatedIn`[47]	Indicates that the subject resource is located in the object feature			`geo:Feature`

Table 8: An example of evaluation entry for the alignment between p_{new} and p_{sw}.

An example of evaluation entry for the alignment between p_{new} and p_{sw} is shown in Table 8.

The following evaluation guidelines have been provided to the raters:

> Express your judgment on the following statement, by assigning one of the values shown in Table 9 to each row.
> *"Properties p_{new} and p_{sw} are interchangeable"*

The results of the user-based evaluation of the alignments between p_{new} and p_{sw} are reported in Table 10. 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).

Likert value	Judgement criteria
Strongly agree	The two properties have exactly the same sense or p_{sw} is more general than p_{new}.
Agree	The two properties have a similar sense, enough to be either used for representing the same relation between subject and object.
Neutral	I am not sure if the two properties have the same sense.
Disagree	The properties have slightly different senses
Strongly Disagree	The properties have completely different senses.

Table 9: Likert scale values and associated criteria for evaluating alignments between p_{new} and p_{sw}

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

6 Discussion and conclusion

The results of our experiment are very promising, although they also show that there is room for improvement.

Non-sense or loose relations. As for the production of p_{new}, the data collected by means of the user-based evaluation is a good source for observing possible emerging patterns indicating when a sentence does not express a strong relation between subject and object of a link. In fact, it can be the case sometimes that a link is put in a sentence just because it is a good practice do to so. Although we have noticed that such cases are not many, detecting them can allow us to improve the precision of our results by removing some of the p_{new} that represent "non-sense" relations.

Better naming. Other emerging patterns can be observed from recurrent subgraph forms connecting subjects and objects, for which we can design more accurate naming strategies for p_{new}, hence improving their cognitive adequacy and alignment results. Additionally, we have analysed only a subset of thematic roles relations that are used by FRED for characterizing the participation of entities in event occurrences. In fact, thematic roles provide useful means for defining labels that can further improve p_{new} naming design strategies.

Alignment to existing Semantic Web properties. As for the alignment procedure, there is also space for improvement, since we addressed this task by computing a simple Levenshtein distance. More sophisticated alignment methods such as those resulting from the Ontology Alignment Initiative[48] or other approaches for entity linking such as SILK[49] [8] 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[50] [4] and LOV[51]. We observed

[48] `http://oaei.ontologymatching.org/`

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

[50] `http://watson.kmi.open.ac.uk/WatsonWUI/`

[51] `http://lov.okfn.org/dataset/lov/`

that properties retrieved from NELL[52] [3] all had an editing distance < 0.70 and almost none of them were judged appropriate. This suggests, in our opinion, that our naming design strategies are closer to the one used by humans, 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 and artificial concatenation of categories learnt automatically.

As for the alignment recall, we could not compute a 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 hypothesize that many cases reveal a lack of intensional coverage in Semantic Web vocabularies, and that our method can help filling this gap.

Conclusion. We have presented a novel approach, named *Legalo*, for uncovering the semantics of hyperlinks based on formal representation of natural language text, and heuristics associated with subgraph-patterns. The main novel aspects of Legalo approach are: property label generation, automatic link tagging, graph-based relation extraction.

Our hypothesis is that pagelinks 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. sentences embedding pagelinks. Our experiment shows promising results: we produce accurate binary properties (F-measure 0.83) and provide alignments with a precision value of 0.84. A demo of Legalo Web service is available online[53], and the binary properties produced in this study can be accessed by means of a sparql endpoint[54].

Ongoing work. Currently, we are developing a generalized version of Legalo, able to work on any webpage and its hyperlinks. Referring to Figure 1, we focus on generalizing component 1*a* and 1*b*, as the *Semantic Web Property Extractor* and the rest of the pipeline are not Wikipedia-specific, although they can be further improved. The challenge is to design strategies for identifying subject and object of a semantic relation denoted by a hyperlink: a task *per-se* on the general Web. A preliminary demo of a generalized approach is available online[55]: it uses NER and disambiguation on DBpedia for addressing subject/object identification. These new developments together with additional evaluation will be included in a possible extended version of this paper[56].

[52] `http://rtw.ml.cmu.edu/rtw/`

[53] `http://wit.istc.cnr.it/legalo/wikipedia`

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

[55] `http://wit.istc.cnr.it/stlab-tools/legalo/`

[56] The last sentence is a note for reviewers that would be removed in a possible final version.

Future work. We intend to perform a large scale processing of hyperlinks for revealing their semantics. Additionally, we want to combine crowd sourcing annotation of results with a learning procedure that would allow Legalo to improve its design strategies, based on emerging recurrent subgraph patterns.

References

1. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
2. J. Bos and M. Nissim. Combining discourse representation theory with framenet. In *Frames, Corpora, and Knowledge Representation*, pages 169–183. R. Rossini Favretti and Bononia University Press, 2008.
3. 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 *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.
4. 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.
5. J. Euzenat and P. Shvaiko. *Ontology Matching, Second Edition.* Springer, 2013.
6. A. Gangemi, A. G. Nuzzolese, V. Presutti, F. Draicchio, A. Musetti, and P. Ciancarini. Automatic typing of dbpedia entities. In *International Semantic Web Conference (1)*, pages 65–81, 2012.
7. Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann, and Martin Bruemmer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW, 5th Wks. on Linked Data on the Web, Lyon, France*, 04 2012.
8. R. Isele and C. Bizer. Active learning of expressive linkage rules using genetic programming. *J. Web Sem.*, 23:2–15, 2013.
9. A. Khalili, S. Auer, and A.-C. Ngonga Ngomo. context lightweight text analytics using linked data. In *Extended Semantic Web Conference (ESWC 2014)*, 2014.
10. Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni. Open language learning for information extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL)*, 2012.
11. G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, Mar. 2001.
12. A. G. Nuzzolese, A. Gangemi, V. Presutti, and P. Ciancarini. Encyclopedic Knowledge Patterns from Wikipedia Links. In L. Aroyo, N. Noy, and C. Welty, editors, *Proceedings fo the 10th International Semantic Web Conference (ISWC2011)*, pages 520–536. Springer, 2011.
13. A. G. Nuzzolese, V. Presutti, A. Gangemi, A. Musetti, and P. Ciancarini. Aemoo: exploring knowledge on the web. In *WebSci*, pages 272–275, 2013.
14. V. Presutti, F. Draicchio, and A. Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In *EKAW: Knowledge Engineering and Knowledge Management that matters (to appear)*. Springer, 2012.