A Semantic Best-Effort Approach for Extracting Structured Discourse Graphs from Wikipedia

André Freitas[1], Danilo S. Carvalho[2], João C. P. da Silva[3], Seán O'Riain[1], and Edward Curry[1]

[1]Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
[2]Department of Systems Engineering and Computer Science (COPPE) &
[3]Computer Science Department
Federal University of Rio de Janeiro (UFRJ)

Abstract. Most information extraction approaches available today have either focused on the extraction of simple relations or in scenarios where data extracted from texts should be normalized into a database schema or ontology. Some relevant information present in natural language texts, however, can be irregular, highly contextualized, with complex semantic dependency relations, poorly structured, and intrinsically ambiguous. These characteristics should also be supported by an information extraction approach. To cope with this scenario, this work introduces a *semantic best-effort information extraction approach*, which targets an information extraction scenario where text information is extracted under a pay-as-you-go data quality perspective, trading high-accuracy, schema consistency and terminological normalization for domain-independency, context capture, wider extraction scope and maximization of the text semantics extraction and representation. A semantic information extraction framework (*Graphia*) is implemented and evaluated over the Wikipedia corpus.

Keywords: Semantic Best-effort extraction, Information Extraction, Semantic Networks, RDF, Linked Data, Semantic Web

1 Introduction

The Linked Data Web brings the vision of a semantic data graph layer on the Web which can improve the ability of users and systems to access and semantically interpret information. Currently most datasets on the Linked Data Web, such as DBpedia, are built from data already structured in different formats, which are mapped to an ontology/vocabulary and are transformed into RDF. Despite its fundamental importance as a grassroots movement to make available a first layer of data on the Web, sharing structured databases on the Web will not be sufficient to make the Semantic Web vision [1] concrete. Most of the information available on the Web today is in a unstructured text format. The integration of this information into the Linked Data Web is a fundamental step towards enabling the Semantic Web vision.

The semantics of unstructured text, however, does not easily fit into structured datasets. While the representation of structured data assumes a high level of regularity, relatively simple conceptual models and a consensual semantics between the users of a structured dataset, the representation of information extracted from texts need to take into account large terminological variation, complex context patterns, fuzzy and conflicting semantics and intrinsically ambiguous sentences. Most information extraction (IE) approaches targeting the extraction of facts from unstructured text have focused on extraction scenarios where accuracy, consistency and a high level of lexical and structural normalization are primary concerns, as in the automatic construction of ontologies and databases. These IE approaches can be complemented by alternative information extraction scenarios where accuracy, consistency and regularity are traded by domain-independency, context capture, wider extraction scope and maximization of the text semantics representation, under a *pay-as-you-go* data quality perspective [8], where data semantics and data quality are built and improved over time. We call an information extraction strategy focused on these aspects a *semantic best-effort information extraction* approach. This type of approach provides a complementary semantic layer, enriching existing datasets and bridging the gap between the Linked Data Web and the Web of Documents.

This work focuses on the construction and analysis of a *semantic best-effort information extraction approach.* The approach extracts *structured discourse graphs* (SDGs) from texts, a representation introduced in [5] which focuses on a RDF compatible graph representation which maximizes the representation of text elements and context under a pay-as-you-go data extraction scenario. Potential applications of this work are: (i) structured and unstructured data integration (ii) open information extraction for IR support, (iii) enrichment of existing Linked Datasets such as DBpedia and YAGO [6].

The contributions of this paper are: (i) deepening the discussion on the pay-as-you-go semantic best-effort information extraction, (ii) a semantic best-effort graph extraction pipeline based on the SDG representation (iii) the implementation of the pipeline in the *Graphia* extraction framework and (iv) the evaluation of the extraction pipeline using Wikipedia as a corpus.

This paper is organized as follows: section 2 provides a motivational scenario based on DBpedia and Wikipedia; section 3 provides a overview of the SDG representation model [5]; section 4 describes the architecture and the components of the semantic best-effort extractor; section 5 provides an experimental analysis of the extraction approach using Wikipedia as a corpus; section 6 analyses the related work in the area; finally, section 7 provides a conclusion and describes future work.

2 Motivational Scenario

The core motivation for a semantic best-effort (SBE) extraction is to provide a structured discourse representation which can enrich datasets with information present in unstructured texts. Currently datasets such as DBpedia are created by extracting (semi-)structured information from Wikipedia. With an appropriate graph representation, it is possible to provide an additional layer for knowledge discovery (KD), search, query and navigation (Figure 1). As a motivational scenario suppose a user wants to know possible connections between Barack Obama and Indonesia. Today this information cannot be directly found in DBpedia, and the user would need to browse and read through Wikipedia articles to find this information. A semantic best-effort structured discourse graph (SDG) can provide an additional link structure extracted from text which, starting from the DBpedia entity Barack Obama, can be used by an application to find the semantic connection with the other DBpedia entity Indonesia. This intermediate layer between text and datasets (Figure 1) has a different level of representation from traditional, ontology-based RDF datasets. In the example, the sentence and its corresponding extracted graph (Figure 1), the temporal references ('from age six to ten') are not resolved to a normalized temporal representation, and only the information present in the verb tense is used to define a temporal context, showing the semantic best-effort/pay-as-you-go nature of the approach. Additionally, the context where the original sentence is embedded in the text is mapped to the graph through a *context_link*. A semantic best-effort extraction/representation provides the core structure of the sentence and its discourse context, maximizing the representation of the text information, allowing the future extension/refinement of the extracted information. The representation of complex and composite relations is a fundamental element in information extraction. In the example scenario, a simple relation extraction would focus on the extraction of triples such as (*Barack Obama, attended, local school*) which does not provide a connection between Barack Obama and Indonesia.

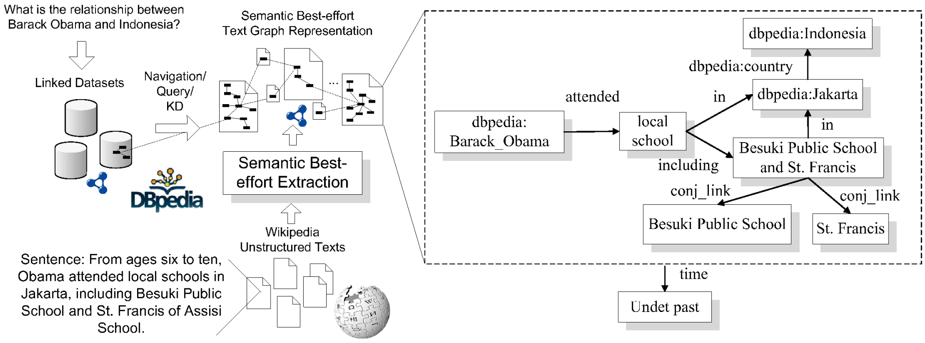
What is the relationship between Barack Obama and Indonesia?
Linked Datasets
Navigation/ Query/ KD
Semantic Best-effort Text Graph Representation
Semantic Best-effort Extraction
Wikipedia Unstructured Texts
Sentence: From ages six to ten, Obama attended local schools in Jakarta, including Besuki Public School and St. Francis of Assisi School.
DBpedia
dbpedia:Indonesia
dbpedia:country
dbpedia:Jakarta
in
attended
dbpedia: Barack_Obama
local school
in
including
Besuki Public School and St. Francis
conj_link
conj_link
Besuki Public School
St. Francis
time
Undet past

Fig. 1. Motivational scenario and example of a SBE graph representation.

3 Representing Text as Discourse Graphs

The objective of structured discourse graphs (SDGs) introduced in [5] is to provide a principled representation for text elements which supports a semantic best-effort extraction. A semantic best-effort (SBE) extraction aims at maximizing the amount of extracted information present in the text, capturing the semantic context and the semantic dependencies where a given fact is embedded. A SBE extraction also minimizes the semantic impact of potential extraction errors by maximizing the semantic isolation between structures associated with different types of extraction operations (e.g. relation extraction, temporal resolution and co-reference resolution) and by facilitating the process of navigating back to the original text source. This isolation facilitates the data consumption/interpretation process under the pay-as-you-go scenario, where the impact of possible incomplete or erroneous extractions is minimized. SDGs provide a representation complementary to Discourse Representation Structures (DRSs). In fact, DRSs can be represented as SDGs [5]. SDGs approach the representation problem from both a data generation (under a SBE scenario) and also from a data consumption perspective. SDGs are designed to be an RDF-based graph representation from the start, also providing a principled semantic interpretation of the graph data through a graph navigation algorithm, facilitating its use under the Linked Data context.

The following items describe the main elements of the structured discourse graph model introduced in [5]. Real sentence graphs extracted from the Wikipedia article *Barack Obama* by the *Graphia*[1] framework are used as examples to introduce the elements of the extraction model. The elements described below are combined into a graph structure which allows a principled algorithmic interpretation model. A more detailed discussion on the SDG representation can be found in [5]. The SDG representation consists of the following core elements:

Named, non-named entities and properties: *Named entities* include categories such as proper nouns, temporal expressions, biological species, substances, among other categories. A named entity is defined by one or more proper nouns (**NNP**) in a noun phrase (**NP**). In RDF, named entities map to instances. *Non-named entities* are more subject to vocabulary variation ('President of the United States', 'American President'), i.e. polysemy and homonymy. Additionally, non-named entities have more complex compositional patterns: commonly non-named entities are composed with less specific named or non-named entities, which can be referenced in different contexts. A non-named entity is defined by one or more nouns (**NN**), adjectives (**JJ**) in a noun phrase (**NP**). In RDF a non-named entity maps to a class which can be referred both as a class and as an instance (*punning*)[2]. *Properties* are built from verbs (**VB**) or from passive verb constructions. Named, non-named entities and properties form the basic triple (relation) pattern which is complemented by the SDG elements below.

Quantifiers & Generic Operators: Represent a special category of nodes which provide an additional qualification over named or non-named entities. Both quantifiers and generic operators are specified by an enumerated set of elements which map to *adverbs*, *numbers*, *comparative and superlative* (suffixes and modifiers). Examples of quantifiers and operators are: *Quantifier:* e.g. one,

[1] http://graphia.dcc.ufrj.br

[2] http://www.w3.org/2007/OWL/wiki/Punning

two, (cardinal numbers), many (much), some, all, thousands of, one of, several, only, most of; *Negation:* e.g. not *Modal:* e.g. could, may, shall, need to, have to, must, maybe, always, possibly; *Comparative:* e.g. largest, smallest, most, largest, smallest. Ex.: Figure 2(E).

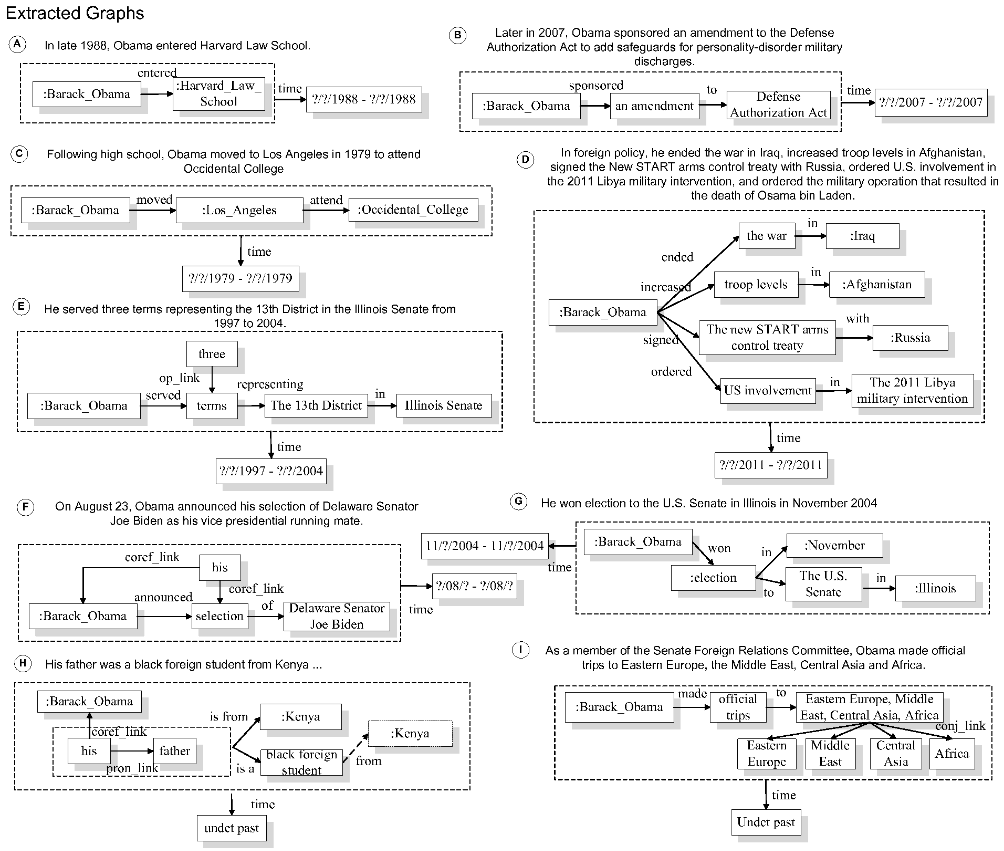
Extracted Graphs
A
In late 1988, Obama entered Harvard Law School.
:Barack_Obama
entered
:Harvard_Law_School
time
?/?/1988 - ?/?/1988
B
Later in 2007, Obama sponsored an amendment to the Defense Authorization Act to add safeguards for personality-disorder military discharges.
:Barack_Obama
sponsored
an amendment
to
Defense Authorization Act
time
?/?/2007 - ?/?/2007
C
Following high school, Obama moved to Los Angeles in 1979 to attend Occidental College
:Barack_Obama
moved
:Los_Angeles
attend
:Occidental_College
time
?/?/1979 - ?/?/1979
D
In foreign policy, he ended the war in Iraq, increased troop levels in Afghanistan, signed the New START arms control treaty with Russia, ordered U.S. involvement in the 2011 Libya military intervention, and ordered the military operation that resulted in the death of Osama bin Laden.
:Barack_Obama
ended
the war
in
:Iraq
increased
troop levels
in
:Afghanistan
signed
The new START arms control treaty
with
:Russia
ordered
US involvement
in
The 2011 Libya military intervention
time
?/?/2011 - ?/?/2011
E
He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004.
three
op_link
:Barack_Obama
served
terms
representing
The 13th District
in
Illinois Senate
time
?/?/1997 - ?/?/2004
F
On August 23, Obama announced his selection of Delaware Senator Joe Biden as his vice presidential running mate.
coref_link
his
coref_link
announced
:Barack_Obama
selection
of
Delaware Senator Joe Biden
?/08/? - ?/08/?
time
G
He won election to the U.S. Senate in Illinois in November 2004
11/?/2004 - 11/?/2004
time
:Barack_Obama
won
:election
in
:November
to
The U.S. Senate
in
:Illinois
H
His father was a black foreign student from Kenya ...
:Barack_Obama
coref_link
his
father
pron_link
is from
:Kenya
is a
black foreign student
from
:Kenya
time
undet past
I
As a member of the Senate Foreign Relations Committee, Obama made official trips to Eastern Europe, the Middle East, Central Asia and Africa.
:Barack_Obama
made
official trips
to
Eastern Europe, Middle East, Central Asia, Africa
conj_link
Eastern Europe
Middle East
Central Asia
Africa
time
Undet past

Fig. 2. Examples of extracted sentence graphs from the Wikipedia article *Barack Obama*. Nodes with a ':' depict entities resolved to DBpedia URIs.

Triple Trees: Not all sentences can be represented in one triple. On a normalized dataset scenario, one semantic statement which demands more than one triple is mapped to a conceptual model structure (as in the case of events for example) which is not explicitly present in the discourse. In the unstructured text graph scenario, sentences which demand more than one triple can be organized into a triple tree. A triple tree is built by a mapping from the syntactic tree of a sentence to a set of triples, where the sentence subject defines the root node of the triple tree. The interpretation of a triple tree is defined by a complete

DFS traversal of the tree, where each connected path from the root node to a non-root node defines an *interpretation path.* Ex.: Figure 2(C).

Context elements: A fact extracted from a natural language text demands a semantic interpretation which may depend on different contexts where the fact is embedded (such as a temporal context). Intra-sentence dependencies are given by dependencies involving a different clause in the same sentence. Intra-sentence context for a triple can be represented by the use of reification (Figure 2). Contexts can also be important to define the semantics of an entity present in two or more triple trees. For example the interpretation of an entity which is neither a root and a leaf node (Figure 2(D)) demands the capture of the pairwise combination of its backwards and forward properties in multiple contexts. This is lost in a typical dereferenciation process where all properties and objects associated with an entity are returned. A third level of context can be defined by mapping the dependencies between extracted triple trees, taking into account the sentences ordering and the relation to text elements in the original discourse. Ex.: Temporal nodes in Figure 2.

Co-Referential elements: Some discourse elements contain indirect references to named entities (*pronominal* & *non-pronominal* co-references). Co-references can refer to either intra or inter sentences named entities. While in some cases co-references can be handled by substituting the co-referent term by the named entity (as in personal pronouns), in other cases this direct substitution can corrupt the semantics of the representation (as in the case of reflexive and personal pronouns) or can mask errors in a semantic best-effort extraction scenario. Co-reference terms include: you, I, someone, there, this, himself, her, this, that, etc. Ex.: Figure 2(F)(H)(I).

Resolved & normalized entities: Resolved entities are entities where a node-substitution in the graph was made from a co-reference to a named entity (e.g. a *personal pronoun* to a named entity). Normalized entities are entities which were transformed to a normalized form. A temporal normalization where date & time references are mapped to a standardized format (September 1st of 2010 mapped to 01/09/2010). Ex.: Figure 2(A)-(G).

4 Structured Discourse Graphs Extraction

4.1 Mapping Natural Language to SDGs

This section describes the basic components of a semantic best-effort extraction pipeline targeting the proposed representation. The extraction pipeline was designed targeting Wikipedia as a corpus. Wikipedia has a *factual discourse*, *a topic-oriented text organization* and *named entities KB given by DBpedia.* The extraction pipeline takes as input Wikipedia texts and returns an extracted RDF graph and a sentence-based graph visualization. The extraction pipeline consists of the following components (Figure 3):

1. Syntactic analysis: The first step in the extraction process is the syntactic parsing of the natural language text into syntactic trees (C-Structures). This module uses the Probabilistic Context-Free Grammar (PCFG) implemented in the Stanford parser. The C-Structures for the sentences are passed to the next modules.

2. Named entity resolution: This component resolves named entities text references to existing DBpedia URIs. The first step consists in the use of the DBpedia Spotlight service[3] where the full article is sent and is returned with annotated URIs. The second step consists in the use of Part-of-Speech tags together with C-Structures to aggregate words into entity candidates which were not resolved by the DBpedia Spotlight service. The entity candidates' strings are sent as search terms to a local entity index which indexes all DBpedia URIs using TF/IDF over labels extracted from the URIs. Returned URIs mapping to the search string terms are used to enrich the original annotated text file with additional URI annotations. The output of this component is the original text with a set of named entity terms annotated with URIs.

3. Personal co-reference resolution and normalization: This component resolves pronominal co-references including personal, possessive and reflexive pronouns. Personal pronouns instances are substituted by the corresponding entities. Possessive and reflexive pronouns are annotated with the corresponding entities that will later define the co-reference links. The co-reference resolution process is done by the pronoun-named entity gender and number agreement (by taking into account gender information present in a name list) and by applying a heuristic strategy based on text distance between the pronoun and named entity candidates. The output of this component are C-Structures with annotated named entities, co-reference substitutions for personal pronouns and possessive and reflexive pronouns annotated with named entities.

4. Graph extraction: The graph extraction module takes as input the annotated C-Structures and generates the triple trees for each sentence by the application of a set of transformation rules based on syntactic conditions through a DFS traversal of the C-Structure. Instead of focusing on terminology-dependent patterns, these rules are based on syntactic patterns. The core set of syntactic rules are split into 6 major categories: *subject*, *predicate*, *object*, *prepositional phrase & noun complement*, *reification*, *time*. Additional details about the graph extraction algorithm can be found online [4] .

1. *Subject: Subjects* are activated by noun phrases (**NP**) when NPs are higher into the syntactic hierarchy and without any NPs as child nodes. This rule applies the following actions: (i) concatenates the nouns in case of compound subjects; (ii) Adds the subject as a node into the triple tree; (iii) adds a URI in case the subject is a named entity.
2. *Predicate: Predicates* are defined by verbal phrases (**VB***). This rule applies the following actions: (i) verifies the verb tense and activates the rule which transforms the verb tense into a temporal representation; (ii) concatenates the neighboring verbs in case there is more than one verb; (iii) verify if the

[3] http://dbpedia.org/spotlight

[4] http://treo.deri.ie/sdg

verb has a property pattern and concatenates the pattern nodes defining them as a labelled edge on the triple tree; (iv) adds the predicate words to the verb/property-pattern and removes these words from the object node in the triple tree; (v) verifies the presence of an explicit temporal reference in the predicate; (vi) Adds the explicit or implicit temporal references as a reification.

3. *Object:* This rule is activated when the search reaches a NP node that does not have a child NP and is after a verb phrase. The rule applies the following actions: (i) identifies the object head; (ii) concatenates the nouns in case of a compound object; (iii) creates an object node with the object in the triple tree; (iv) in case the words in the node correspond to a recognized entity, adds the associated URI.
4. *Prepositional phrase & Noun complement:* This rule is activated when the search finds a NP node that does not have a NP as a child and that has a prepositional phrase (**PP**) as a sibling node. The goal of this rule is to find ownership relations in subjects and objects. The rule applies the following actions: (i) concatenates the words in the noun phrase; (ii) creates a graph node connected by an edge with a preposition.
5. *Reification:* This rule is activated when the search finds a preposition node. It ignores the prepositional phrases which modifies NPs, which are handled by the previous rule. The rule applies the following actions: (i) concatenates the words in the PP, excluding the preposition; (ii) creates a reification for the prepositional phrase; (iii) verifies the existence of explicit temporal references and creates temporal reification nodes.
6. *Time:* This rule is not applied over the nodes of the syntactic structure and it is indirectly invoked by the other rules. This rule identifies explicit and implicit date references. Dates are detected by a set of regular expressions, which detects and normalizes explicit date references to a predefined format. Implicit date references (verb tenses) are detected by the analysis of POS tags.

5. Graph construction: This component receives the triple trees from the previous component and outputs the final graph serialization. Context URIs are created among different sentences and among each sentence and the article context URI. Additionally, local URIs are created for each resource which was not resolved to a DBpedia URI.

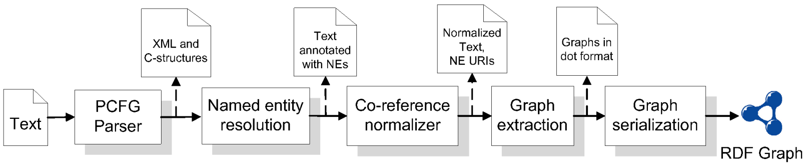

Fig. 3. High-level architecture of the SBE graph extraction pipeline.

5 Extraction & Evaluation

This section focuses on the analysis of the feasibility of a semantic best-effort extraction by evaluating the proposed extraction pipeline. The *key questions* that are targeted by the evaluation are: (i) the verification of the feasibility of extracting structured discourse graphs following the SDG representation; (ii) the quantification of the errors associated with each extraction step and (iii) the determination of which extraction error mostly impacts the semantics of the extracted graph.

The evaluation methodology is based on the work of Harrington & Clark [2], which selected a list of sample factual articles associated with named entities, and evaluated the extracted semantic networks according to a set of errors. The evaluation differs in relation to the corpus (here the corpus is the English Wikipedia) and on the final set of error categories (the error categories in this work target the generation of the core elements of the representation). Articles were selected randomly satisfying the following criteria: 2 articles about people, 2 articles about organizations and 1 article about a place. Each article has a number of characters greater than 40K. The article size served as an indicator of a more diverse discourse sample base and of the quality of the discourse. The selected articles were: Apple Inc., Google, Napoleon, Paris, John Paul II.

Error Categories	Apple	Google	Napoleon	John Paul	Paris	Avg.
Reification construction	0.20	0.13	0.10	0.06	0.17	**0.132**
Pronominal co-reference	0.13	0.10	0.03	0.00	0.03	**0.058**
Conjunction	0.00	0.06	0.03	0.03	0.06	**0.036**
Named entity	0.06	0.06	0.00	0.06	0.10	**0.056**
Subject construction	0.10	0.06	0.10	0.10	0.20	**0.112**
Object construction	0.16	0.23	0.26	0.23	0.34	**0.244**
Triple tree construction	0.33	0.26	0.20	0.30	0.31	**0.280**
Predicate construction	0.23	0.23	0.06	0.13	0.03	**0.136**
Explicit temporal reference	0.16	0.03	0.10	0.20	0.06	**0.110**
Accuracy						
Correct graphs	0.39	0.46	0.40	0.56	0.43	**0.448**
Complete graphs	0.16	0.23	0.20	0.16	0.06	**0.162**
Interpretable graphs	0.99	0.96	0.93	0.96	0.94	**0.956**

Table 1. Accuracy and frequency of extraction error categories.

The quality of the extraction was manually evaluated for each graph generated from a sentence. Sentences which were not well-formed or which were classified as outside the scope of the extraction pipeline (sentences with complex *subordination* structures [4]) were removed from the evaluation set. The final dataset consists of 1033 relations (triples) from 150 sentences which were manually classified [4] . Comparatively, for a related work using human-based evaluation, Harrington & Clark [2] evaluates approximately 160 relations and 5 topics. The extraction pipeline was implemented in Python following the architecture outlined in the previous section. A web evaluation platform was built to allow an efficient manual evaluation process. In the evaluation platform the original natural language sentence and a visualization of the extracted graph are displayed to a human evaluator, who classifies the sentences in relation to: (i) 10 sentence features (to guarantee an heterogeneous and complete sample set, which evaluates all aspects of the extraction pipeline), (ii) 9 error categories (indicate the quality impact of each pipeline component) and (iii) the accuracy of the extraction (to evaluate how each error category impacts the final extraction). The list of sentence features can be found online [4] . Table 1 shows the categorized frequency of errors for each article together with the associated extraction accuracy. To evaluate the accuracy in a semantic best-effort scenario three measures were defined: the *correctness*, the *completeness* and the *interpretability* of the graph extractions. These three measures represent different levels of accuracy: A *correct graph* is an extracted graph which is fully consistent with the semantic model; a *complete graph* is a correct graph which maps all the information of a sentence, and an *interpretable graph* is a graph fragment which has the correct semantics of its *basic triple paths* (core *s, p, o* pattern from the main clause), despite the possible presence of extraction errors in other extracted structures (such as co-reference links and reified statements). The correctness of the basic triple paths is the most important element in the extraction, highly impacting the interpretability and usability of the extracted SDG.

The high percentage of interpretable graphs, shows that there is a basic triple path which is correct in **95.6%** of the extracted graphs. The extractor is able to extract an informational and correct fragment in practically all the sentences. **55.2%** of graphs contained some extraction error. Only **16.2%** of the extracted graphs mapped all the information contained in the sentence, which shows the major direction for improvement (completeness), but which is aligned with a pay-as-you-go scenario. The major justification for the lack of completeness is the fact that the SBE extractor, in many occasions, ignores sentence structures which are not central (e.g. appositive) and do a partial extraction. The most impacting error categories were triple trees, object and reification construction, categories which are strongly interrelated. The error frequencies indicate that the existing extractor still needs to be improved in relation to object construction criteria, in particular in relation to the extraction of non-named entities. The low frequency of errors related to named and temporal entities shows the robustness on the determination of these semantic pivots. The relatively high reification construction error frequency shows that the breadth of the rules for extracting prepositional phrases is still limited. The proposed representation supported the best-effort extraction by isolating errors from different parts of the extraction pipeline, keeping a high number of graph fragments interpretable even when a component of the pipeline fails. The final extracted graphs were easily represented as RDF. However, the centrality on the modelling of context brings mechanisms such as reifications, named graphs (quads) and quints to the center of the discussion for text representation.

6 Related Work

Existing related work can be classified in three main categories: *semantic networks extraction from texts* [2, 3], *open relation extraction* [7, 4] and *ontology extraction from Wikipedia* [6].

Harrington & Clark [2] describe AskNet, an information extraction system which builds large scale semantic networks from unstructured texts. The extraction pipeline of AskNet starts with the parsing of text sentences using the C&C parser [2], a parser based on the linguistic formalism of Combinatory Categorial Grammar (CCG). A Named Entity Recognition (NER) stage is performed using the C&C NER tagger. After the sentences are parsed, AskNet uses the Boxer semantic analysis tool [2], which produces a first-order logic representation based on the semantic model of the Discourse Representation Theory (DRT). A low coverage pronoun resolution approach is used for pronominal co-references. Wojtinnek et al. [3] provides an introductory discussion on the RDF translation of the AskNet output. No principled discussion on the discourse and graph representation is provided in [2, 3]. Despite having similar objectives, the approach used in the SBE extraction pipeline is significantly different (parser, NER and pronominal co-reference resolution strategy). On the representation side, this work targets a graph representation and algorithmic interpretation which focuses on RDF and is not directly mediated by DRT.

TextRunner [7] is an open information extraction (domain independent) framework. TextRunner uses a single-pass extractor consisting of a POS-tagger and a lightweight noun phrase chunker to determine the core entities in a sentence, normalizing relations by removing less semantically significative terms (e.g. modifiers), defining a probabilistic redundancy model based on the frequency of normalized facts as a correctness estimator. Comparatively, TextRunner focuses on the extraction of simple relations and does not cover the representation of more complex discourse structures. Co-reference resolution is not covered in its extraction process. Nguyen et al. [4] propose an approach for relation extraction over Wikipedia by mining frequent subsequences from the syntactic and semantic path between entity pairs in the corpus. The approach uses dependency structures and semantic role labelling and does not focus on the extraction and representation of complex relations.

YAGO2 is an extension of YAGO which targets the extraction and representation of temporal and spatial statements. To assign a spatio-temporal dimension to the facts, a new representation (SPOTL(X)) is proposed. The focus on Wikipedia, the centrality of the representation of reifications, and the definition of a temporal model are common aspects between YAGO2 and this work.

7 Conclusion & Future Work

This work focuses on the analysis of a semantic best-effort extraction approach using structured discourse graphs (SDGs), a RDF-based discourse representation format. A semantic best-effort extraction pipeline is proposed and is implemented on the *Graphia* framework. The quality of the proposed extraction approach is evaluated over Wikipedia. The final extraction achieved **44.8%** of correctness, **16.2%** completeness and an interpretability of **95.6%**. The representation played a key role in isolating errors from different components of the extraction pipeline, impacting on the interpretability performance. The final approach showed a high coverage of the elements of the SDG representation model, with the evaluation pointing into a main direction for improvement: increasing the extraction completeness. The evaluation of error categories shows that this can be achieved by improving non-named entity recognition criteria and the treatment of prepositional phrases.

Acknowledgments. The work presented in this paper has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. T. Berners-Lee, J. Hendler and O. Lassila, The Semantic Web, Scientific American, May, p. 29-37, 2001.
2. B. Harrington and S. Clark. ASKNet: Creating and Evaluating Large Scale Integrated Semantic Networks. Intl. Journal of Semantic Computing, 2(3), 2009.
3. P-R. Wojtinnek, B. Harrington, S. Rudolph and S. Pulman. Conceptual Knowledge Acquisition Using Automatically Generated Large-Scale Semantic Networks. In Proc. of the 18th Intl. Conference on Conceptual Structures, 2010.
4. D. P.T Nguyen, Y. Matsuo and M. Ishizuka, Exploiting Syntactic and Semantic Information for Relation Extraction from Wikipedia. IJCAI Workshop on Text-Mining & Link-Analysis, 2007.
5. A. Freitas, D.S. Carvallho, J.C. P. da Silva, S. O'Riain, E. Curry. A Structured Discourse Graph Representation for a Semantic Best-Effort Text Extraction (to appear), available at: http://treo.deri.ie/sdg, 2012.
6. J. Hoffart, F. Suchanek, K. Berberich and G. Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Special Issue of the Artificial Intelligence Journal, 2012.
7. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead and O. Etzion. Open Information Extraction from the Web. In Proc. of the Intl. Joint Conference in Artificial Intelligence, 2007.
8. M. J. Franklin, A. Y. Halevy, D. Maier: From databases to dataspaces: a new abstraction for information management. SIGMOD Record 34(4): 27-33, 2005.