Extracting Events from Wikipedia as RDF Triples Linked to Widespread Semantic Web Datasets

Carlo Aliprandi[1], Francesco Ronzano[2], Andrea Marchetti[2], Maurizio Tesconi[2], and Salvatore Minutoli[2]

[1] Synthema Srl
Via Malasoma 24
56121 Ospedaletto (Pisa) - Italy
carlo.aliprandi@synthema.it
[2] Institute of Informatics and Telematics (IIT) CNR
Via G. Moruzzi, 1
56123 Pisa - Italy
{francesco.ronzano,andrea.marchetti,maurizio.tesconi,
salvatore.minutoli}@iit.cnr.it

Abstract. Many attempts have been made to extract structured data from Web resources, exposing them as RDF triples and interlinking them with other RDF datasets: in this way it is possible to create clouds of highly integrated Semantic Web data collections. In this paper we describe an approach to enhance the extraction of semantic contents from unstructured textual documents, in particular considering Wikipedia articles and focusing on event mining. Starting from the deep parsing of a set of English Wikipedia articles, we produce a semantic annotation compliant with the Knowledge Annotation Format (KAF). We extract events from the KAF semantic annotation and then we structure each event as a set of RDF triples linked to both DBpedia and WordNet. We point out examples of automatically mined events, providing some general evaluation of how our approach may discover new events and link them to existing contents.

Keywords: Knowledge Representation, Knowledge Extraction, Semantic Web, Natural Language Processing, Semantics.

1 Introduction

The core aim of the Semantic Web is to provide a set of methodologies, standards, technologies and best practices to make explicit the semantics that lies behind the data exposed over the Web. As a consequence, it is possible to support an easy and serendipitous automatic integration of the great variety of Web contents, thanks also to the exploitation of shared knowledge references like ontologies, lexicons and semantic resources. In this scenario, the Resource Description Framework (RDF) [1] and the Ontology Web Language (OWL) [2] currently constitute the two core W3C standards useful to respectively represent knowledge over the Web and to specify a formalized semantic reference frame.

A.A. Ozok and P. Zaphiris (Eds.): Online Communities, HCII 2011, LNCS 6778, pp. 90–99, 2011.

Many methodologies to extract knowledge from Web contents have been proposed. They can be divided into two wide groups:

- *relation or fact extraction systems*: they usually apply specific Web mining procedures to extract facts from online contents. Facts are usually represented as attribute-value pairs describing some feature of a given entity of interest, i.e. the population of a specific country. Extracted facts are usually shown to users as search results referring to the part of Web documents. Examples are WebKnox [3], the Grazer System [4] and TextRunner [5].
- *interlinking systems:* they mine information from poorly structured Web contents and, unlike the previous group of systems, they expose the extracted knowledge over the Web as RDF data. In this context, the Linked Data initiative [6] has defined a set of best practices to represent such knowledge by exploiting RDF and to unambiguously identify entities on a Web scale by means of URIs. A central role is played by DBpedia [7], an extract from Wikipedia contents, representing a hub for many other Linked Data datasets.

Many other proposals to produce Semantic Web interlinked datasets have been made in parallel to Linked Data. They usually deal with some particular kind of Web contents and propose specific methodologies, like:

- *systems to enrich Web content:* they apply procedures for keyword extraction or Named Entity recognition over Web pages, automatically producing a set of relevant terms to be annotated through the URI of the referred Wikipedia/DBpedia entity. Open Calais [8], for example, parses documents and points out entities, facts and events. When possible, entities are linked to DBpedia, Freebase or GeoNames URIs. Wikify [9] performs keyword extraction from Web pages, and disambiguates mined terms linking them to the referred Wikipedia entity.
- *systems to enrich social tagging service:* systems like Faviki [10] and LODr [11] allow users to associate a tag to the Wikipedia page describing the referred concept. A different approach to automatically connect user tags to a specific concept of Wikipedia has been adopted by [12], exploiting also Tagpedia [13], a semantic resource for Tag Sense Disambiguation.

Considering the attempts to build semantic resources by mining Wikipedia, [14] describes a set of methodologies adopted to extract from Wikipedia an association thesaurus, by exploiting the set of internal links, the taxonomy of Wikipedia categories and by mining the contents of each article. [15] describes how to build Tagpedia, a semantic reference useful to support the disambiguation of tags, by mining the structure of Wikipedia articles.

Proposals to extract text snippets from Wikipedia representing facts have been also defined. Specific extraction techniques have been tuned to gather relevant text snippets from other articles [16] or to mine Wikipedia Named Entities over time [17].

Mining of Wikipedia has also been carried out by applying Natural Language Processing: a dump of the English Wikipedia has been shallow parsed and semantically annotated [18]. Applying both shallow and deep parsing to Wikipedia, methodologies to build a common sense knowledge base have been proposed [19].

In this paper we propose an approach to mine the unstructured textual data of Wikipedia so as to extract events, representing them through sets of RDF triples and integrating them in the Linked Data dataset. In section 2, after having clarified the notion of event, we present the KYOTO Annotation Format (KAF) and we describe Synthema English Slot Grammar, a deep parser used to automatically produce KAF annotations. In section 3 we describe the procedure to identify events and to represent them through RDF triples exploiting also WordNet and DBpedia. In section 4 we provide some meaningful examples of this procedure as well as some initial evaluation. We conclude in section 5, discussing our future plans to improve the event extraction process from KAF annotated texts.

2 Annotating Textual Documents by Exploiting KAF

The goal of this paper is to explore the possibility to mine events from linguistically and semantically annotated textual documents and to try to propose possible ways to represent them as sets of RDF triples tightly connected with core Semantic Web datasets like DBpedia. For this reason we need first of all to define what we mean by an event, but also to specify a possible RDF representation of events as useful Semantic Web knowledge.

We assert that *an event is something that happens having some relevance in providing information in a particular context: it could be characterized by specific spatial and temporal coordinates*. As a consequence, an event is usually built around a specific action or happening.

In order to mine events from documents and, in particular, from the contents of Wikipedia articles, we consider the results of their linguistic and semantic analysis encoded in KAF [20], the deep semantic annotation format that we developed in the context of the KYOTO Project. KAF is a language neutral annotation format representing both morpho-syntactic and semantic annotation of documents through a layered structure. Starting from the lower of all its annotation layers, where tokens, sentences and paragraph are identified, in KAF each additional layer is built on top of the lower one, referring to its constituent elements. In this way, several levels of text annotation can be added by different linguistic processors. Moreover, specialized linguistic processors can be developed to generate incremental annotations for each specific layer.

In KAF there are three macro-layers of document annotation (see also Figure 1):

- *morpho-syntactic layer*: it groups all the language-specific text annotations. Tokens, sentences and paragraphs are identified in a specific document. Terms made of words or multi-words are pointed out, along with their Part Of Speech. In this layer also chunks and functional dependencies are represented.
- *level-1 semantic layer*: it includes linear annotation of expressions of time, events, quantities and locations.
- *level-2 semantic layer*: it is mainly devoted to represent facts, in a non linear annotation context, thus possibly aggregating evidences from the lower layers of multiple textual sources.

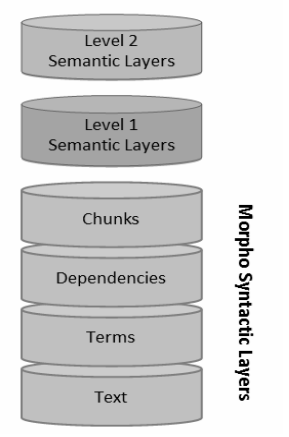

Fig. 1. The three macro-layers of KAF document annotation

Our RDF event extraction process mainly exploits the results of the morpho-syntactic annotation layer of KAF. In this layer, the following elements are annotated:

- *word forms*: each word form is unambiguously determined and linked to the sentence and paragraph of the text it belongs to.
- *terms*: terms, also composed of two or more word forms, are identified and characterized by lemma, Part Of Speech and, when possible, by the type of referred Named Entity. The link of a term to an external reference can be represented. This is the KAF feature that we have mostly exploited to represent term meaning, either as a WordNet concept, disambiguated by proper WSD, or as a DBpedia entity.
- *dependency relations*: they represent functional relations among terms, such as *Agent*, *Action*, *Object*, *Qualifier*, *When*, *Where*, *How*.
- *chunks*: they are used to identify structured phrases, spanning one or more terms, like noun phrases, verbal phrases and prepositional phrases.

Synthema English Slot Grammar (Syn ESG), an inverse parallel deep parser, is used to automatically produce KAF annotations. It carries out complex Natural Language Processing (NLP) tasks like word tokenization, segmentation, Part Of Speech tagging, dependency parsing, anaphora resolution and functional analysis.

Syn ESG is intended to identify relevant knowledge from the raw text, by detecting concepts and semantic relations in texts. Concept extraction and text mining are applied through a pipeline of linguistic and semantic processors that share as a common ground McCord's theory of Slot Grammar [21]. Syn ESG parser - a bottom-up chart parser - employs a parse evaluation scheme used for pruning away unlikely analyses during parsing as well as for ranking final analyses. By including semantic information directly in the dependency grammar structures, the system relies on the lexical semantic information combined with functional relations. Besides Named Entities, locations, time-points and dates, Syn ESG detects relevant information like

chunks, noun phrases and verbal phrases. The detected terms are then extracted, reduced to their Part Of Speech and functional tagged base form.

Finally, syntactically and semantically tagged words are properly encoded in the corresponding KAF annotation, and the specific KAF layer is produced.

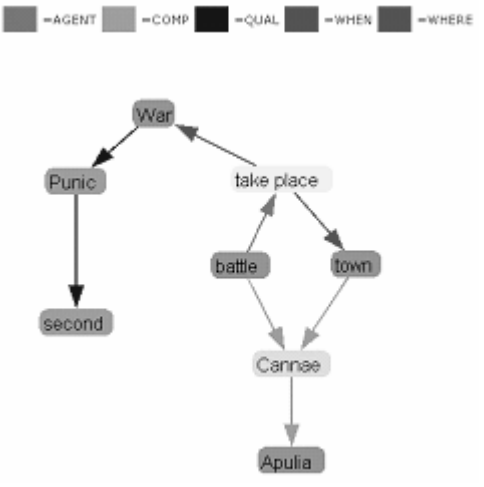

Fig. 2. Example output produced by the Syn ESG text parser

In Figure 2 we show the output sample for the sentence: "*The Battle of Cannae took place near the town of Cannae in Apulia, during the Second Punic War*", taken from the Wikipedia article describing the *Battle of Cannae*. The figure shows functional dependencies among terms. Note the reference resolution for the word "*Cannae*", that is correctly co-referred to a unique URI (represented by the id #2583).

Starting from the described KAF features, in the following sections we detail the extraction of events from KAF annotated documents as well as their representation as RDF triples.

3 Events Extraction and RDF Representation

As mentioned before, in a KAF annotated text, mined terms are linked, thanks to a specific Word Sense Disambiguation algorithm, to either WordNet or DBpedia. Summarizing, we assert that *the meaning of a term in KAF documents can be denoted by a synsetID of WordNet and/or by a URI of an entity of DBpedia*. Unlike DBpedia URIs, WordNet synsetIDs may represent ambiguous Web identifiers. In this paper we are proposing a tentative URI schema for WordNet synsets: the URI that identifies a synset of English WordNet version 3.0 is '*http://www.kyoto-project.eu/wordnet/English/30/synsetID*'.

In the next section we describe how to extract events from KAF documents and how to represent them as sets of RDF triples.

3.1 Identifying Events Inside KAF Annotated Documents

To extract RDF representations of events from a KAF annotated document, we mainly consider the terms linked to WordNet synsets and to DBpedia entities, and the dependency relations linking terms.

We base our event extraction process on the assumption that the nucleus of an event is an action expressed by a verb, i.e. by a term that has been classified as a verb and associated to a WordNet synset. We also assume that the features that characterize an event are included in and represented by the set of dependency relations of the sentence the verb belongs to and connected in some way to the same verb.

In particular we consider the dependency relations of *subject* (the entity that performs the action), *object* (the entity affected by the action), *when* (time or interval of happening of the action) and *where* (place of happening of the action).

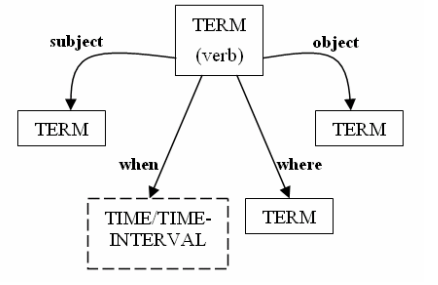

Fig. 3. Set of dependency relations characterizing an event

To identify an event we need a set of dependency relations, in particular we need at least a *subject* or an *object* dependency relation. The connection of terms with WordNet or DBpedia URIs is essential in order to represent events as Semantic Web RDF triples linked with other datasets. In Figure 3 the set of dependency relations that can be exploited to characterize an event is schematized. Each term is connected to WordNet or DBpedia except the time-interval characterizing the *when* dependency relation, which is a literal.

Given the previous set of dependency relations, in the next section we describe how to represent events as a set of RDF triples.

3.2 Representing Events as RDF Triples Exploiting DBpedia URIs

Each event that we identify in a document is characterized by a subset of the dependency relations shown in Figure 3. In addition, the verbal WordNet synset defining the action of the event identified by its synset URI and the terms of the *subject*, *object* and *where* dependency relations are identified by a URI of a WordNet synset or of a DBpedia entity. The *when* dependency relation is described by a string specifying a time or interval.

We can express an event as a set of RDF triples by exploiting ontological properties taken from the DBpedia Ontology or from a properly structured ontology of dependency relations. In what follows we assume use of an ontology, referred

to as Dependency Relation Ontology, properly published at the stub '*http://dependencyrelat.org/ontology/*' namespace. The *when* and *where* dependency relations can be respectively expressed by the '*http://dbpedia.org/ontology/Date*' and '*http://dbpedia.org/ontology/Place*' OWL properties of the DBpedia Ontology. The *subject* and *object* dependency relations can be respectively represented in the Dependency Relation Ontology by the '*http://dependencyrelat.org/ontology/Subject*' and '*http://dependencyrelat.org/ontology/Object*' properties.

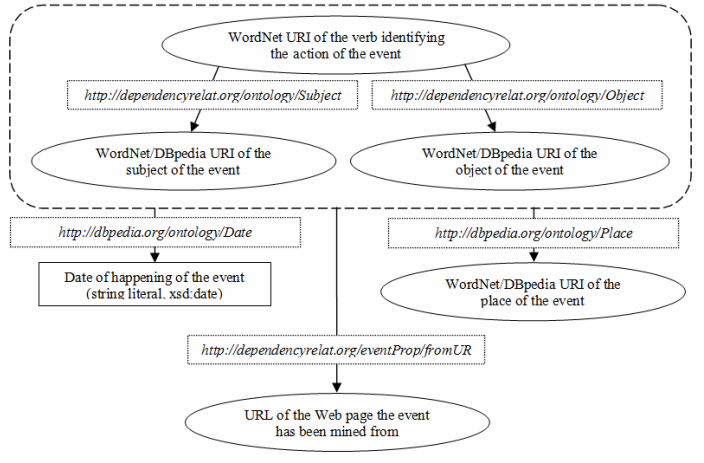

Fig. 4. RDF representation of an event

As a consequence, we obtain the RDF representation of an event shown in Figure 4. The two triples describing the subject and the object of an event are reified and thus referenceable by a URI: we assume to use for this purpose URIs published at the stub '*http://linkedevents.org/events/*' namespace. The two URIs describing these RDF triples are in turn grouped under the same RDF Bag Container and thus both referenceable through a third URI that is also assumed to be published at the same stub namespace. This URI, referred to as the *event URI*, obtained by aggregating in the same RDF Bag Container the URIs defined reifying the subject and object triples, is represented in Figure 4 by the dashed line. It points out the core features of an event and represents the subject of the RDF triples describing the place and time of happening of that event (properties '*http://dbpedia.org/ontology/Place*' and '*http://dbpedia.org/ontology/Date*').

To preserve in the RDF representation, also the URL of the Web page the event has been mined from, we also represent the *event URI* by means of the stub '*http://dependencyrelat.org/eventProp/fromURL*' ontological property.

To conclude, we can notice that the two RDF triples pointing out the *subject* and the *object* of an event could be collapsed into a single one if we manage to represent WordNet verbs, constituting the core actions characterizing an event, as ontological properties: we can exploit the hypernym/hyponym taxonomy of WordNet verbs to define a hierarchy of subsumed ontological properties related to them (i.e. in English WordNet 3.0 the verb *consume* subsumes the verb *eat*). Each of these properties

contributes to the definition of the main RDF triple of an event by linking the WordNet/DBpedia URI of the subject of the action to the WordNet/DBpedia URI that points out the object, in this case through the URI of the ontological properties representing the verb describing the actions.

4 Example of Mined Events

To gather initial examples of facts, we annotated a set of Wikipedia articles, mainly related to wars and battles. Here we show a significant example of mined events from a sentence extracted from the Wikipedia page describing the 'Battle of Gettysburg':

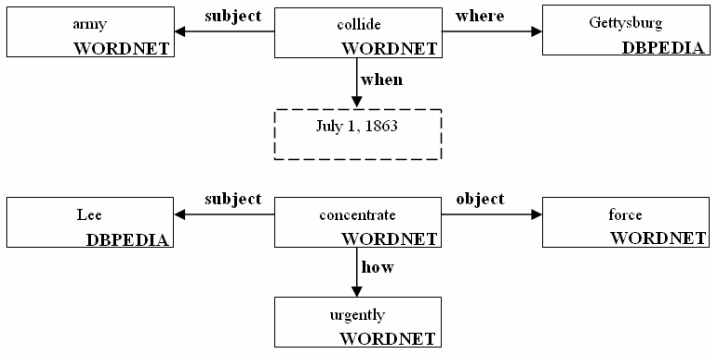

Fig. 5. Concepts mined from the sample sentence: "The two armies began to collide at Gettysburg on July 1, 1863, as Lee urgently concentrated his forces there"

Figure 5 shows terms and dependency relations mined from the sentence. Two sentences are identified: "*The two armies began to collide at Gettysburg on July 1, 1863"* and *"Lee urgently concentrated his forces there."*. Moreover, five terms are connected to WordNet synsets and two terms are identified through DBpedia entities. The date is a literal and thus managed as a string (July 1, 1863).

From this sentence we can mine two events; their RDF representation is shown in Figure 6. Notice that two DBpedia URIs are exploited to refer to the two terms of the sentence:

1. Gettysburg: http:// dbpedia.org/resource /Gettysburg,_Pennsylvania
2. Robert E. Lee: http:// dbpedia.org/resource /Robert_E._Lee

Moreover, the URI of four WordNet Synsets are exploited to represent the RDF triples of these events. In particular, the event related to the first sentence is described by a subject- The corresponding RDF triple is reified and further characterized by the place and date of happening of that event (Gettysburg, July 1st 1863). The event related to the second sentence is characterized by a subject and an object but no place and date are specified. Both events are linked to the URL of the Wikipedia page they have been mined from, through the RDF property: '*http://dependencyrelat.org/eventProp/fromURL*'.

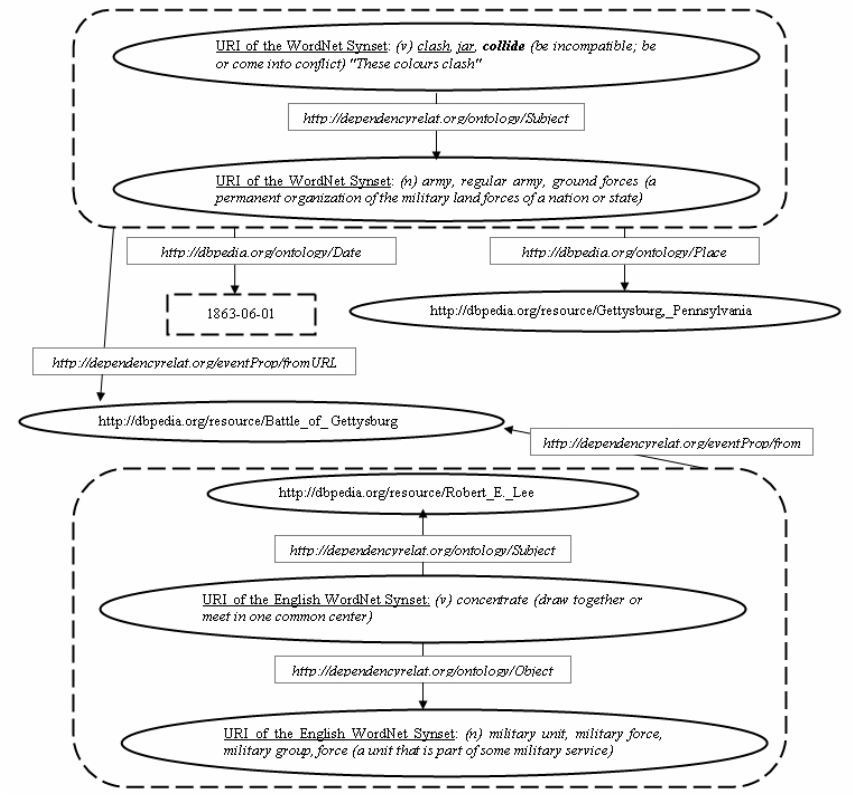

Fig. 6. RDF representation of mined events from the sample sentence

5 Conclusions and Future Work

In this paper we have presented an approach to extract events from documents, annotated using a deep text parser according to the KYOTO Annotation Format.

We have defined an event extraction methodology that takes as input KAF annotated documents: it is based on functional dependencies and on the results of disambiguation of the terms through WordNet synsets or DBpedia entities. We have defined a representation of the mined events, as a set of RDF triples exploiting the URI of WordNet and DBpedia to point out the different entities taking part into events. We have provided a significant example of event mining procedure starting from a Wikipedia article. We are carrying out a global evaluation of the effectiveness of our event mining procedure by processing a larger set of Wikipedia articles. In this way we plan to provide also global quantitative evaluations of the quality of events.

Even though the event extraction activities are at an early stage, we believe that the methodologies described can be useful to automatically produce Semantic Web data linked with other datasets. In particular, by applying this process to DBpedia, a considerable amount of new RDF triples can be generated. In conclusion, we can state

that our approach represents an example of synergy between NLP techniques and the Semantic Web to produce and link over the Web semantically-grounded contents.

Acknowledgments. This work is partially funded by the European Commission (KYOTO project, ICT-2007-211423).

References

1. RDF W3C Web Page, http://www.w3.org/RDF/
2. OWL W3C Recomm., http://www.w3.org/TR/owl-features/
3. Urbansky, D., Thom, J.A.: WebKnox: Web Knowledge Extraction. In: 13th Australasian Document Computing Symposium, Hobart (2008)
4. Zhao, S., Betx, J.: Corroborate and Learn Facts from the Web. In: 13th International Conference on Knowledge Discovery and Data Mining, San Josè (2007)
5. Banko, M., Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction. In: 46th ACL: Human Language Technologies, Columbus (2008)
6. Linked Data Web Site, http://linkeddata.org/
7. DBpedia Web Site, http://dbpedia.org/About
8. Open Calais Web Site, http://www.opencalais.com/
9. Wikify! Web Site, http://www.wikifyer.com/
10. Faviki Web Site, http://www.faviki.com/
11. Passant, A.: LODr - A Linking Open Data Tagging System. In: Social Data on the Web Workshop at the 7th Int. Semantic Web Conference, Karlsrhue (2008)
12. Tesconi, M., Ronzano, F., Marchetti, A., Minutoli, S.: Semantify del.icio.us: automatically turn your tags into senses. In: Social Data on the Web Workshop at the 7th International Semantic Web Conference, Karlsrhue (2008)
13. Tagpedia Web Site, http://www.tagpedia.org/
14. Nakayama, K.: Extracting Structured Knowledge for Semantic Web by Mining Wikipedia. In: Social Data on the Web Workshop at the 7th International Semantic Web Conference, Karlsrhue (2008)
15. Ronzano, F., Marchetti, A., Tesconi, M., Minutoli, S.: Tagpedia: a Semantic Reference to Describe and Search for Web Resources. In: Social Web and Knowledge Management Workshop at the 17th World Wide Web Conference, WWW 2008, Beijing (2008)
16. Adafre, S.F., Jijkoun, V., de Rijke, M.: Fact Discovery in Wikipedia. In: IEEE/WIC/ACM International Conference on Web Intelligence, Silicon Valley (2007)
17. Bhole, A., Fortuna, B., Grobelnik, M., Mladenic, D.: Mining Wikipedia and Relating Named Entities over Time. In: 13th International Conference on Knowledge Discovery and Data Mining, San Josè (2007)
18. Asterias, J., Zaragoza, H., Ciaramita, M., Attardi, G.: Semantically Annotated Snapshot of the English Wikipedia. In: 6th International Language Resources and Evaluation Conference LREC 2008, Marrakech (2008)
19. Suh, S., Halpin, H., Klein, E.: Extracting Common Sense Knowledge from Wikipedia. In: 6th International Semantic Web Conference, Athens, GA, USA (2006)
20. Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Aliprandi, C., Monachini, M.: KAF: a generic semantic annotation format. In: 5th International Conference on Generative Approaches to the Lexicon, Pisa (2009)
21. McCord, M.C.: Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars. Natural Language and Logic, 118–145 (1989)