# A Comparison of Knowledge Extraction Tools for the Semantic Web

Aldo Gangemi[1,2]

[1] LIPN, Université Paris13-CNRS-SorbonneCité, France
[2] STLab, ISTC-CNR, Rome, Italy.

**Abstract.** In the last years, basic NLP tasks: NER, WSD, relation extraction, etc. have been configured for Semantic Web tasks including ontology learning, linked data population, entity resolution, NL querying to linked data, etc. Some assessment of the state of art of existing Knowledge Extraction (KE) tools when applied to the Semantic Web is then desirable. In this paper we describe a landscape analysis of several tools, either conceived specifically for KE on the Semantic Web, or adaptable to it, or even acting as aggregators of extracted data from other tools. Our aim is to assess the currently available capabilities against a rich palette of ontology design constructs, focusing specifically on the actual semantic reusability of KE output.

## 1 Introduction

We present a landscape analysis of the current tools for Knowledge Extraction from text (KE), when applied on the Semantic Web (SW).

Knowledge Extraction from text has become a key semantic technology, and has become key to the Semantic Web as well (see. e.g. [31]). Indeed, interest in ontology learning is not new (see e.g. [23], which dates back to 2001, and [10]), and an advanced tool like Text2Onto [11] was set up already in 2005.

However, interest in KE was initially limited in the SW community, which preferred to concentrate on manual design of ontologies as a seal of quality. Things started changing after the linked data bootstrapping provided by DBpedia [22], and the consequent need for substantial population of knowledge bases, schema induction from data, natural language access to structured data, and in general all applications that make joint exploitation of structured and unstructured content. In practice, also Natural Language Processing (NLP) research started using SW resources as background knowledge, and incrementally graph-based methods are entering the toolbox of semantic technologies in the large.

As a result, several tools have appeared providing useful, scalable, application-ready and precise learning of basic semantic data structures, such as tagged named entities, factual relations, topics for document classification, and their integration with SW languages is growing fast. These tools are the bulk of the set considered in this study.

On the other hand, the SW community soon realized that learning just basic semantic data structures fails to achieve complex analytic KE tasks that require e.g. event recognition, event dependency detection, logical relation induction,

etc. For example [5] points against the topological sparsity of the results of early ontology learning even at the schema (TBox) level (let alone at the data level), and proves the importance of reusing ontology patterns for improving the topological connectedness of learnt ontologies.

Very recently, more tools are appearing that attempt a deeper KE, typically by hybridizing statistical (trained models) and rule-based methods, and taking advantage of existing knowledge from Linked Open Data as well as of smart heuristics that cling to all sorts of features and structures that become incrementally available on the Web. These tools are also considered in this study.

This study does not intend to be complete in terms of tools tested, parameters singled out for testing, or sample size used in testing. On the contrary, as a *landscape analysis*, it aims to indicate the problems encountered, and some directions and solutions, in order to prepare the ground for a substantial benchmark and a reference evaluation procedure for KE tools on the SW (KE2SW tools).

In Section 2 we make a short recap of the efforts in abridging linguistic and formal semantics, which is the central problem of KE2SW. In Section 3 we survey parameters that can be applied to the comparison between tools for KE2SW: tool performance, structural measures, basic tasks across NLP and SW applications. In Section 4 we describe the text used in the comparison, and the testing principles. In Section 5 we describe the tools. In Section 6 we present the measures obtained from running the tools on the test text, and discuss them.

## 2 Knowledge extraction and the Semantic Web

Traditionally, NLP tasks are distinguished into basic (e.g. named entity recognition), and applied (e.g. question answering). When we try to reuse NLP algorithms for the SW, we can also distinguish between basic (e.g. class induction) and application tasks (NL querying of linked data). In this landscape analysis, we map NLP basic tasks to SW ones, and compare different tools with respect to possible functionalities that accomplish those tasks.

The semantics provided by NLP resources is quite different from that assumed for ontologies in knowledge representation and the SW in particular. Moreover, with the exception of formal deep parsing, e.g. based on Discourse Representation Theory (DRT) [21], or Markov Logic [13], the (formal) semantics of NLP data is fairly shallow, being limited to intensional relations between (multi-)words, senses, or synsets, informal identity relation in entity resolution techniques, sense tagging from typically small sets of tags (e.g. WordNets "super senses"), lightweight concept taxonomies, etc.

The actual exploitation and enrichment of ontologies partly relies on the ability to reuse NLP results after appropriate conversion. Such ability is exemplified in some academic and industrial applications that label these techniques as "semantic technology". The current situation of semantic technology can be summarized as in Figure 1, which depicts the relations between formal and linguistic knowledge: linguistic knowledge uses formal background knowledge, but can enable access to formal knowledge (and enrich it) as well. The union of for-

mal and formalized linguistic knowledge can be further extended by means of automated inferences.

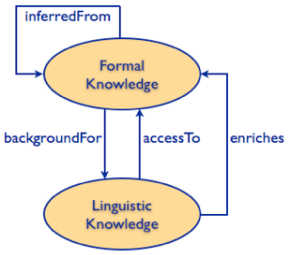


Fig. 1: Hybridization of formal and linguistic knowledge in semantic technologies.

Despite recent uprise in adoption of NLP techniques for SW and conversely of SW knowledge for NLP, there is still a large gap between the data structures of lexical and NLP data, and the formal semantics largely adopted for ontologies in the Semantic Web. Current proposals of schemas and formats for abridging NLP and SW, e.g. LMF [15], SKOS-XL [25], LIR [32] Lemon [24],[3] FISE[4], NIF [19], with implementations like Apache Stanbol[5] and NERD[6] are helpful, but they address primarily the "porting" or "lifting" of NLP results or lexical resources to the SW, while the problem of formally reusing NLP results in the SW is mostly left to the choice of specific applications or users. It is therefore interesting to assess the current situation at the tool level, in order to look at the possible best practices that are emerging, as well as to stress them a little bit in order to figure out what can be done in practice, even when there is no direct abridging between the two worlds.

## 3 Parameters

As we stated in the introduction, this study has no pretense of completeness over all the tools that can be used for KE on the SW: we have tested some of them with a setting that is an attempt to clarify the actual functionalities available when we make KE for a SW application, and we have to figure out the formal semantics of the extracted structures. In other words, the major contribution of the study is a clarification of *what* we do when we use KE for the SW, with an explicit intention to map linguistic semantics into formal semantics. A more complete survey is planned for a journal version of this study.

We firstly distinguished among measures addressing system-level features (time performance, export capabilities, standard compliance), structural measures of the produced ontologies (axioms, density, presence of linguistic annotations, textual grounding, etc.), and measures of achievements with reference to *basic tasks*. Only the third type of measurements has been carried out in this study. The measures precision $p$, recall $r$ and accuracy $a$ have been applied (when possible) to a subset of the following parameters related to *basic tasks*, with correspondence between NLP and SW terminology, which of course reflects the different notions of semantics usually assumed in the two fields of expertise:[7]

[3] The W3C Ontology-Lexicon Community Group (`http://www.w3.org/community/ontolex/wiki/Main_Page`) is active on drafting a standard out of Lemon.

[4] `http://stanbol.apache.org/docs/trunk/components/enhancer/enhancementstructure.html#fisetextannotation`

[5] `http://dev.iks-project.eu:8081/enhancer`

[6] `http://nerd.eurecom.fr`

[7] Many tasks have quite a large literature, and we can hardly summarize it here; reference work is cited for some of them.

1. topic extraction (recognition of specific topics, e.g. individuals from the range of the property `dc:subject`), see also [3]
2. named entity recognition (individual induction) [27]
3. named entity resolution (identity resolution for individuals) [4]
4. named entity coreference (coreference of individuals) [29]
5. terminology extraction (induction of constants pertinent to a domain, typically for classes or properties) [18]
   (a) class induction
   (b) property induction
6. sense tagging (≈ class membership induction) [8]
7. sense disambiguation (≈ identity resolution for classes) [28]
8. taxonomy induction (≈ subclass relation induction) [33]
9. (non-taxonomic, non-role, binary) relation extraction (property assertion –fact– induction) [9,2]
10. semantic role labeling (≈ property induction for events and n-ary relations) [26]
11. event detection (≈ n-ary relationship induction) [20]
12. frame detection (≈ n-ary relation –type– induction) [12]

There are other basic tasks that have not been tested, because some are mostly unknown to NLP, some only have approximate counterparts in knowledge representation for the SW, some have been noticed during the study, but are not well attested in either literature. These include at least: schema-level logical structure extraction [35,6,13]: class equivalence, union of classes, class covering, class disjointness, disjoint partition, restriction induction (that in NLP is part of e.g. automatic formalization of glosses); as well as data-level logical structure extraction [6,13]: entity linking (identity/difference between individuals), individual conjunction (complex object), individual disjunction (collection), fact negation (negative property assertion), factual entailment (≈ dependency relation between events or reified relationships) [1], etc.

In order to give a more formal basis to the correspondences provided in the previous list, we have reconstructed some of the current translation practices from NLP to formal semantics, and reported them in Table 1. By no means these are definitive recommendations for translation, due to the variety of requirements and domains of application, which can motivate different choices. For the tasks that have been tested, when RDF or OWL representation of extraction results is not provided by the tools, we have applied Table 1 as a set of default assumptions for translation.[8]

## 4 The sample text

The sample used in this study has been taken from an online article of The New York Times[9] entitled "Syrian Rebels Tied to Al Qaeda Play Key Role in War",

[8] With conjunction or disjunction of individuals, an ontology can be used to represent e.g. *collections* and their members, or *complex entities* and their parts.

[9] `http://www.nytimes.com/2012/12/09/world/middleeast/syrian-rebels-tied-to-al-qaeda-play-key-role-in-war.html`

| | |
|---|---|
| Topic | `<Document> dc:subject <Topic>` |
| Named entity | `owl:NamedIndividual` |
| Entity resolution (NE) | `owl:sameAs` |
| Entity coreference | `owl:sameAs` |
| Term | `owl:Class \|\| owl:ObjectProperty \|\| owl:DatatypeProperty` |
| Sense tag | `owl:NamedIndividual rdf:type owl:Class` |
| Sense disambiguation (classes) | `owl:equivalentClass` |
| Taxonomy (subclasses) | `owl:subClassOf` |
| Extracted (binary) relation | `owl:ObjectProperty \|\| owl:DatatypeProperty` |
| Semantic role | `owl:ObjectProperty \|\| owl:DatatypeProperty` |
| Event | `<Event> rdf:type <Event.type> . <Event> <semrole`$_i$`> <Entity`$_j$`>` |
| Frame | `<Event.type> owl:subClassOf <Frame>` |
| Restriction | `owl:Restriction` |
| Linked entities | `owl:sameAs \|\| owl:differentFrom` |
| Conjunct of individuals | `owl:NamedIndividual` |
| Disjunction of individuals | `owl:NamedIndividual` |
| Factual entailment | `<Event`$_1$`> <dependency> <Event`$_2$`>` |

Table 1: Translation table, used when default assumptions are to be applied on the results of a tool. The output of basic tasks not listed here is trivially translated according to model-theoretic semantics (e.g. "union of classes").

and its size has been cut to 1491 characters in order to adapt it to the smallest maximum size of texts accepted by the tested tools (Section 5).[10] The text is cited here (minor typographic editing has been performed for character encoding compatibility across the tools):

*The lone Syrian rebel group with an explicit stamp of approval from Al Qaeda has become one of the uprising most effective fighting forces, posing a stark challenge to the United States and other countries that want to support the rebels but not Islamic extremists. Money flows to the group, the Nusra Front, from like-minded donors abroad. Its fighters, a small minority of the rebels, have the boldness and skill to storm fortified positions and lead other battalions to capture military bases and oil fields. As their successes mount, they gather more weapons and attract more fighters. The group is a direct offshoot of Al Qaeda in Iraq, Iraqi officials and former Iraqi insurgents say, which has contributed veteran fighters and weapons. "This is just a simple way of returning the favor to our Syrian brothers that fought with us on the lands of Iraq," said a veteran of Al Qaeda in Iraq, who said he helped lead the Nusra Front's efforts in Syria. The United States, sensing that time may be running out for Syria president Bashar al-Assad, hopes to isolate the group to prevent it from inheriting Syria or fighting on after Mr. Assad's fall to pursue its goal of an Islamic state. As the United States pushes the Syrian opposition to organize a viable alternative government, it plans to blacklist the Nusra Front as a terrorist organization, making it illegal for Americans to have financial dealings with the group and prompting similar sanctions from Europe.*

We have produced one ontology from the output of each tool from the list in Section 5, translating it when necessary according to the default assumptions as in Table 1, or editing it when RDF or OWL parsing was difficult.

[10] One text only may seem a small sample even for a landscape analysis, but in practice we had to measure 14 tools across 15 dimensions, with a total amount of 1069 extracted constructs, 727 of which are included in the merged ontology, and 524 in the reference ontology.

As explained in Section 3, we want to assess some information measures on the produced ontologies, and we need some reference knowledge space for that. We have chosen the simplest way to create such a knowledge space: a reference ontology. But how to produce it without introducing subjective biases or arbitrary design decisions?

For this study we have decided not to produce a "gold standard" ontology from a top-down, intellectual ontology design interpreting the text. This choice is due to a lack of requirements: ontology design and semantic technologies are highly dependent on application tasks and expert requirements: it would be too subjective or even unfair to produce an ontology based on an average or ideal task/requirement set.

A possible solution is to choose specific application requirements, and to design the ontology based on them, e.g. "find all events involving a terroristic organization". Another solution is to "merge" all the results of the tested tools, so that each tool is *comparatively* evaluated within the semantic tool space. Of course, the merged ontology needs to be cleaned up of all errors and noise coming from specific tools, in order to produce a reference ontology. This solution is inspired by the typical testing used in information retrieval with incomplete information [7], where supervised relevant results from different methods are merged in order to provide a baseline.

The second solution seemed more attractive to us because if makes us free from the problem of choosing a task that does not look like biasing the test towards a certain tool. It is also interesting as an indicator of how far "merging tools" like Apache Stanbol or NERD can be pushed when integrating multiple KE outputs[11].

The produced ontologies, including the merged and the reference ones, are available online.[12] An analysis of the results based on the measures listed in Section 3 is reported in Section 6.

## 5 Tools

The tools considered share certain characteristics that make them a low hanging fruit for our landscape analysis. They are available as easily installable downloadable code, web applications, or APIs, and at least in public demo form. They are also tools for Open Domain information extraction, which means that they are not dependent on training to a specific domain[13]. Their licensing has not been investigated for this study, because we are interested in assessing the state of art functionalities, rather than their legal exploitability in either commercial or academic projects. We have not confined our study to tools that can produce SW output (typically RDF or OWL), because it is usual practice to reuse KE output in SW tools. Therefore, in cases where RDF or OWL is not produced by the tool, we have applied default assumptions on how to convert the output (see

[11] In the planned extended survey, we will consider also other experimental settings, including explicit requirements, user behavior, etc.

[12] `http://stlab.istc.cnr.it/documents/testing/ke2swontologies.zip`

[13] This is not totally true for PoolParty Knowledge Extractor, but its dependency is harmless for the sake of this study.

Section 3). Finally, certain tools can be configured (in terms of confidence or algorithm to be used) in order to optimize their performance: in this study, we have sticked to default configurations, even if this choice might have penalized some tools (in particular Apache Stanbol).

The following tools have been selected:

- AIDA[14] is a framework and online tool for named entity recognition and resolution. Given a natural-language text or a Web table, it maps mentions of ambiguous names onto canonical entities (e.g., individual people or places) registered in the YAGO2 knowledge base[15], used also to provide sense tagging. AIDA can be configured for the algorithm to be applied (prior probability, key phrase similarity, coherence). It is available as a demo web application or as a Java RMI web service [36].
- AlchemyAPI[16] uses machine learning and natural language parsing technology for analyzing web or text-based content for named entity extraction, sense tagging, as well as for relationships and topics. It does not provide a direct RDF encoding. It is available as a demo web application or as a REST service, also for mobile SDKs.
- Apache Stanbol[17] is an Open Source HTTP service meant to help Content Management System developers to semi-automatically enhance unstructured content with semantic annotations to be able to link documents with related entities and topics. Current enhancers include RDF encoding of results from multilingual named entity recognition and resolution, sense tagging with reference to DBpedia and GeoNames, text span grounding, confidence, and related images. It is available as a demo web application, as a REST service, or downloadable.
- DBpedia Spotlight[18] is a tool for automatically annotating mentions of DBpedia resources in text. It is available as a demo web application, as a REST service, or downloadable.
- CiceroLite[19] (formerly known as Extractiv), performs named entity recognition for English, Arabic, Chinese, and a number of European-language texts. It also performs sense tagging, relation extraction, and semantic role labeling. It is available as a demo web application, and as a REST service.
- FOX[20] is a merger and orchestrator of KE tools, focusing on results that include named entity recognition and resolution, sense tagging, term extraction, and relation extraction. It provides an ontology that generalizes over the sense tags provided by the merged tools. FOX also uses NIF [19] to generalize over textual grounding methods;[21]. It is available as a demo web application.
- FRED[22] is a tool for automatically producing RDF/OWL ontologies and linked data from text. The method is based on deep semantic parsing as implemented Boxer [6], Discourse Representation Theory [21], Linguistic Frames [30], and Ontology Design Patterns [16]. Results are enriched with NER from the Semiosearch Wikifier (see below). It is available as a demo web application, as a REST service,

[14] http://www.mpi-inf.mpg.de/yago-naga/aida/

[15] http://www.mpi-inf.mpg.de/yago-naga/yago

[16] http://www.alchemyapi.com/api/demo.html

[17] http://dev.iks-project.eu:8081/enhancer

[18] http://dbpedia-spotlight.github.com/demo

[19] http://demo.languagecomputer.com/cicerolite

[20] http://aksw.org/Projects/FOX.html

[21] http://ontowiki.net/Projects/FOX/files?get=fox_evaluation.pdf

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

or downloadable. The current output of FRED is either graphic or in Turtle encoding: the second is an "intermediate" specification, which is typically refactored in order to comply to the type of text analyzed: encyclopedic definitions, factual information, etc. [34]

- NERD[23] [17] is a merger of KE tools (at the time of writing: AlchemyAPI, DBpedia Spotlight, Extractiv, Lupedia, OpenCalais, Saplo, SemiTags, Wikimeta, Yahoo! Content Analysis, and Zemanta), currently focusing on results that include named entity recognition and resolution, and sense tagging. It provides a simple ontology that generalizes over the sense tags provided by the merged tools. NERD also uses NIF [19]. It is available as a demo web application, and as a web service, with APIs for Java and Python.
- Open Calais[24] is a KE tool that extracts named entities with sense tags, facts and events. It is available as a web application and as a web service. It has been used via the web application for homogeneity with the other tools. We have also tried the Open Calais TopBraid Composer[25] plugin, which produces an RDF file automatically. The RDF schemata used by Open Calais have a mixed semantics, and have to be refactored in order to be used as a formal output that is relevant to the domain addressed by the text.
- PoolParty Knowledge Discoverer[26] is a text mining and entity extraction tool based on knowledge models, thesauri and linked data. Content, categories, images and tags are recommended automatically when controlled vocabularies are used as a base knowledge model. In other words, Knowledge Discoverer is dependent on a reference knowledge base typically derived from some controlled vocabularies, e.g. a thesaurus. Configuring one controlled vocabulary instead of another makes results completely different. for our test, we have checked it with two configurations: "all kind of topics", and "economy". It is available as a demo web application.
- ReVerb[27] is a program that automatically identifies and extracts binary relationships from English sentences. ReVerb is designed for web-scale information extraction, where the target relations cannot be specified in advance. ReVerb runs on a model trained out of the big dataset of Open Information Extraction web triples. ReVerb takes raw text as input, and outputs (argument1, relation phrase, argument2) triples. It can be downloaded and there is a related web application[28], not used for this study because it does not accept bulk text [14].
- Semiosearch Wikifier[29] resolves arbitrary named entities or terms (i.e. either individuals or concepts) on DBpedia entities by integrating several components: a named entity recognizer (currently Alchemy[30]), a semiotically informed index of Wikipedia pages (text is selected from page sections and metadata according to explicit formal queries), as well as matching and heuristic strategies. It is available as a demo web application.
- Wikimeta[31] is a tool for multilingual named entity recognition and resolution, and sense tagging. It links texts data to concepts of the Linked Open Data network

[23] http://nerd.eurecom.fr

[24] http://viewer.opencalais.com/

[25] http://www.topquadrant.com/products/TB_Composer.html

[26] http://poolparty.biz/demozone/general

[27] http://reverb.cs.washington.edu

[28] http://openie.cs.washington.edu/

[29] http://wit.istc.cnr.it/stlab-tools/wikifier

[30] http://www.alchemyapi.com/api/demo.html

[31] http://www.wikimeta.com/wapi/semtag.pl

through various sources like DBpedia, Geonames, CIA World Factbook or directly to Wikipedia or the web when there is no available resource. It is available as a demo web application and as a REST service.
– Zemanta[32] provides enriched content for articles, images and websites to bloggers. It matches text with publicly available content and displays it in the creation tool as it is being written. Behind its interaction capabilities, it does named entity recognition and resolution, as well as content linking. It is available as a demo web application and as an API for content management systems.

## 6 Results and discussion

We firstly include (Table 2) a table including all the tools with their featured tasks. We have considered only a subset of the basic tasks (1 to 12), from the list given in Section 3. Some measures of these 12 tasks are not included in the paper for space reasons, but are available online[33]. Tool-level and structural measures have not been addressed in this landscape analysis. We have made an assessment

| Tool | Topics | NER | NE-RS | TE | TE-RS | Senses | Tax | Rel | Roles | Events | Frames |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AIDA | – | + | + | – | – | + | – | – | – | – | – |
| Alchemy | + | + | – | + | – | + | – | + | – | – | – |
| Apache Stanbol | – | + | + | – | – | + | – | – | – | – | – |
| CiceroLite | – | + | + | + | + | + | – | + | + | + | + |
| DB Spotlight | – | + | + | – | – | + | – | – | – | – | – |
| FOX | + | + | + | + | + | + | – | – | – | – | – |
| FRED | – | + | + | + | + | + | + | + | + | + | + |
| NERD | – | + | + | – | – | + | – | – | – | – | – |
| Open Calais | + | + | – | – | – | + | – | – | – | + | – |
| PoolParty KD | + | – | – | – | – | – | – | – | – | – | – |
| ReVerb | – | – | – | – | – | – | – | + | – | – | – |
| Semiosearch | – | – | + | – | + | – | – | – | – | – | – |
| Wikimeta | – | + | – | + | + | + | – | – | – | – | – |
| Zemanta | – | + | – | – | – | – | – | – | – | – | – |

Table 2: Summary of featured basic tasks (as obtained from testing)

of the precision, recall, F-measure, and accuracy of the tools distinguishing them by basic tasks. Measures have been calculated on the merged ontology for each one of the 11 tasks, so that the merged output is used as the upper limit for the measurement. Only five measures are included in the paper (see infra for a site where all measures can be browsed).

Topic extraction tools produce output including broad topics (Alchemy and Open Calais), topics resolved into Wikipedia categories (PoolParty), subject tags (Alchemy), and social tags (Open Calais). We have decided to treat them all as topics, since a real distinction is very hard to make at the theoretical level, while the methods to extract them (e.g. from social tag spaces or Wikipedia) are relevant for the specific task, but do not impact much at the level of produced

[32] http://www.zemanta.com/demo/

[33] http://stlab.istc.cnr.it/stlab/KnowledgeExtractionToolEval

| *Topic Ex Tool* | *p* | *r* | $F_1$ | *a* |
|---|---|---|---|---|
| Alchemy | .74 | .50 | .60 | .52 |
| OpenCalais | 1.00 | .28 | .44 | .48 |
| PoolParty KE | .50 | .22 | .30 | .28 |

Table 3: Comparison of topic extraction tools.

| *Sense Tagging Tool* | *p* | *r* | *F*1 | *a* |
|---|---|---|---|---|
| AIDA | 1.00 | .57 | .73 | .64 |
| Alchemy | 1.00 | .57 | .73 | .64 |
| Apache Stanbol | 1.00 | .43 | .60 | .50 |
| CiceroLite | .64 | .64 | .64 | .54 |
| DBpedia Spotlight | .83 | .36 | .50 | .42 |
| FOX | 1.00 | .50 | .67 | .57 |
| FRED+SST | .75 | .43 | .55 | .48 |
| NERD | .90 | .64 | .75 | .69 |
| OpenCalais | 1.00 | .50 | .67 | .57 |
| Wikimeta | .85 | .79 | .81 | .80 |
| Zemanta | 1.00 | .21 | .35 | .27 |

Table 4: Comparison of sense tagging tools.

knowledge, unless there is a resolution performed with respect to e.g. Linked Open Data (this is true only for PoolParty Knowledge Discoverer). Table 3 contains the results, and show very different performances. 64 topics have been extracted and merged by the three tools, with an overall precision (manually evaluated after merging) of .72.

Named entity recognition in this KE2SW study was assessed only for named entities that are typically represented as individuals in an ontology, while the named entities that are typically appropriate to class or property names are assessed in the terminology extraction and resolution measures (not presented here). After merging and cleaning, 58 named entities remained for evaluation. Table 5 contains the results for this task, showing here a quite consistent behavior across tools. Out of the 58 named entities (individuals) extracted and merged, the overall precision (manually evaluated after merging) is .25. Alchemy, AIDA, and Zemanta stand out on all measures.

Several issues have been encountered when merging and cleaning the results from the different tools. In some cases, named entities have been given directly in terms of *resolved* entities: we have decided to evaluate them as correct or wrong based on the validity of the resolution, even if there is no specific indication of the phrase that has been recognized. In some cases, terms have been recognized instead of named entities: when these are actually referential usages of terms (e.g. "the rebels") they have been accepted as individuals, otherwise they counted as errors. Finally, we had to decide if we need to count tokens (multiple references to the same entity in text) or just types. After a detailed scrutiny, the effect of tokens on precision and recall seemed negligible (two failed recognitions added by tokens across all tools), so we decided to skip tokens for this study.

Table 6 contains the results for the named entity resolution task. Out of the 19 named entities (individuals) that have been resolved, the overall precision (manually evaluated after merging) is .55. AIDA stands out in terms of precision and accuracy, while Wikimeta is high on recall. Most resolutions are made with respect to DBpedia entities.

| *NER Tool* | *p* | *r* | *F*1 | *a* |
|---|---|---|---|---|
| AIDA | 1.00 | .57 | .73 | .89 |
| Alchemy | 1.00 | .57 | .73 | .89 |
| Apache Stanbol | .55 | .43 | .48 | .77 |
| CiceroLite | .79 | .79 | .79 | .89 |
| DBpedia Spotlight | .75 | .21 | .33 | .79 |
| FOX | .88 | .50 | .64 | .86 |
| FRED | .73 | .57 | .64 | .84 |
| NERD | .73 | .79 | .76 | .88 |
| Open Calais | .70 | .50 | .58 | .82 |
| Wikimeta | .71 | .71 | .71 | .86 |
| Zemanta | .92 | .79 | .85 | .93 |

Table 5: Comparison of named entity recognition tools.

| *NE Resolution Tool* | *p* | *r* | *F*1 | *a* |
|---|---|---|---|---|
| AIDA | 1.00 | .64 | .78 | .80 |
| Apache Stanbol | .33 | .36 | .35 | .25 |
| CiceroLite | 1.00 | .55 | .71 | .75 |
| DBpedia Spotlight | .75 | .27 | .40 | .55 |
| FOX | .88 | .64 | .74 | .75 |
| FRED+Semiosearch | .80 | .36 | .50 | .60 |
| NERD | 1.00 | .27 | .43 | .60 |
| Semiosearch | .67 | .55 | .60 | .60 |
| Wikimeta | .71 | .91 | .80 | .75 |

Table 6: Comparison of named entity resolution tools.

| *Term Extraction Tool* | *p* | *r* | *F*1 | *a* |
|---|---|---|---|---|
| Alchemy | .76 | .16 | .26 | .20 |
| CiceroLite | 1.00 | .17 | .29 | .21 |
| FOX | .90 | .27 | .42 | .33 |
| FRED | .93 | .89 | .91 | .90 |
| Wikimeta | 1.00 | .03 | .06 | .04 |

Table 7: Comparison of terminology extraction tools.

| *Term Resolution Tool* | *p* | *r* | *F*1 | *a* |
|---|---|---|---|---|
| CiceroLite | 1.00 | .05 | .10 | .07 |
| FOX | .71 | .63 | .67 | .65 |
| FRED+Semiosearch | 1.00 | .05 | .10 | .07 |
| Semiosearch | .41 | .47 | .44 | .46 |
| Wikimeta | .33 | .05 | .09 | .07 |

Table 8: Comparison of terminology resolution tools.

Table 4 contains the results for the sense tagging task. 19 named entities have been tagged, and the type triples have been merged, with an overall precision (manually evaluated after merging) of .74. Overall, the tools performed quite well on this task (with Wikimeta standing out on recall and accuracy), confirming the good results from literature when using DBpedia and other linked data as background knowledge.

Table 7 contains the results of the terminology extraction task. 109 terms have been extracted and merged by five tools, with an overall precision (manually evaluated after merging) of .94, and with FRED standing out on all measures. Table 8 contains the results of "terminology resolution", which is typically the output of a Word Sense Disambiguation (WSD) algorithm; however, the tested tools do not include any WSD components, therefore disambiguation is just the result of performing NE resolution on terms that refer to classes rather than to individuals. Indeed, only 35 out of 109 terms (.32%) have been resolved, with an overall precision of .54. FOX stands out in this task, with an accuracy of .65.

Table 9 contains the results for the relation extraction task. The variety of relations found is here very high, since the techniques and the assumptions on the relation pattern to discover are very different. In particular, FRED is based on neo-Davidsonian event-reification semantics, for example it represents the sentence: *they gather more weapons* as an event `gather_1` with the semantic role: `agent` and `theme`, played by the entities `thing_1` and `weapon_1`. On the contrary, Alchemy and ReVerb follow a strict binary style, e.g. they ex-

| *RelEx Tool* | *p* | *r* | *F*1 | *a* |
|---|---|---|---|---|
| Alchemy | .69 | .25 | .37 | .30 |
| CiceroLite | .90 | .20 | .33 | .25 |
| FRED | .84 | .82 | .83 | .82 |
| ReVerb | .67 | .23 | .34 | .27 |

Table 9: Comparison of relation extraction tools.

| *Event Detection Tool* | *p* | *r* | *F*1 | *a* |
|---|---|---|---|---|
| CiceroLite | 1.00 | .14 | .24 | .18 |
| FRED | .73 | .93 | .82 | .87 |
| Open Calais | .50 | .03 | .06 | .04 |

Table 10: Comparison of event detection tools.

tract a relationship `gather(they,more weapons)`. CiceroLite has an intermediate approach, trying to guess the arity of the relation, and here it has a binary: `gather(they,weapons)`.

In the previous example, it seems quite efficient to go with the binary style, because the relation/event is used with two explicit arguments. However, things change when there are more than two arguments. For example, with the sentence: *it plans to blacklist the Nusra Front as a terrorist organization*, binary-style tools do not go very far. There are important cohesion aspects here that are hardly caught by means of simple triple patterns: *it* is an anaphora for `United States`, `blacklist` is used with three explicit arguments, and `plan` is used with two, but one of them is the sub-sentence governed by `blacklist`. Here are the representations given by the four tools in this case:

- (ReVerb): no extraction
- (Alchemy): `plans to blacklist(it,the Nusra Front as a terrorist organization)`
- (CiceroLite): `plans to blacklist(it,front,(AS) a terrorist organization)`
- (FRED): `experiencer(plan_3,United_States) ; theme(plan_3,blacklist_3) ; agent(blacklist_3,United_States) ; patient(blacklist_3,NusraFront) ; as(blacklist_3,organization_3) ; TerroristOrganization(organization_3)`

For this task, we have decided to exclude event reification, which is instead tested as a separate task. However, this choice does not penalize FRED, because besides semantic roles, it infers "semantic-web-style" binary relations. For example, from the phrase: *its fighters*, where *its* is an anaphora to Al-Qaeda, FRED extracts: `fighterOf(fighter_1,AlQaeda)`.

When merging the results, we have then considered only non-role binary relations from the four tools, generating 62 relations, with an overall precision (manually evaluated after merging) of .71. The results for this task seem then very promising, and deserve further investigation on how much integration can be done among the different perspectives. For example, a stronger merging could be done by mapping reified events from FRED or CiceroLite to purely binary relations from Alchemy or ReVerb. This may be done in OWL2 by exploiting punning constructs.

Table 10 contains the results for the event detection task. Only 3 tools contain such functionality: CiceroLite, FRED, and Open Calais. As we commented in the previous test, in order to perform event detection, a tool needs also to perform semantic role labeling. FRED and Open Calais also apply some typing of events and values filling the roles, so that they can also be considered "frame detection" tools [34]. For example, Open Calais provides the following frame on top of a detected event from the sentence: *the United States pushes the Syrian opposition to*

*organize a viable alternative government*: `DiplomaticRelations(diplomaticentity: United States ; diplomaticaction: opposition ; diplomaticentity: viable alternative government).`

FRED is the only tool here that provides RDF output (at least from the web applications that we have tested), and resolves event frames onto reference lexicons (VerbNet and FrameNet). After merging the results, we have generated 40 events, with an overall precision (manually evaluated after merging) of .73. The difference in recall is meaningful in this task (as well as in the previous one): FRED uses a categorial parser to extract syntactic structures that are formalized as events (i.e. it provides *deep parsing*, while the other tools apparently use a purely statistical approach with *shallow parsing*, which is known to reach a much lower recall on this task. FRED stands out also in precision, which seems to confirm that deep parsing approach positively correlates with good results on relation and event extraction.

The measures on semantic role labeling and frame detection (only available on FRED and CiceroLite) are not shown here for space reasons[34], but they contain a detailed analysis of the elements extracted for the task: semantic roles, correctness of roles, correctness of fillers, correctness of frames, and coreference resolution. If we simply sum all elements (297 in total), FRED performed better, with an accuracy of .82.

## 7 Conclusions

We have presented the results of a landscape analysis in the area of knowledge extraction for the Semantic Web (KE2SW). We have investigated the feasibility of a comparison among KE tools when used for SW tasks. We have proved that this is feasible, but we need to create formally correct correspondences between NLP basic tasks, and SW population basic tasks. Design activities to obtain semantic homogeneity across tool outputs is required. In addition, evaluation and measures differ across different tasks, and a lot of tool-specific issues emerge when comparing the outputs. This study results to be a first step in the creation of adequate benchmarks for KE applied to SW, and proves the importance of integrating measurement of different tasks in the perspective of providing useful analytic data out of text. Future work includes an actual experiment on a larger dataset, also exploiting integration functionalities provided by platforms like NERD, FOX and Stanbol.

A practical conclusion of this study is that tools for KE provide good results for all the tested basic tasks, and there is room for applications that integrate NLP results for the Semantic Web. Firstly, the measures for merged ontologies result to be good enough, and we imagine optimization methods to filter out the contribution coming from worst tools for a certain task. Secondly, with appropriate semantic recipes (transformation patterns), the production of merged ontologies can be automatized. Merging and orchestrating applications like Stanbol Enhancers, NERD and FOX with standards like NIF, are on the right track, and refactoring components like Stanbol Rules[35] make it possible to customize the output in appropriate ways to reasoning over the Web the Data.

[34] They are available at `http://stlab.istc.cnr.it/stlab/SRLFE`

[35] `http://stanbol.apache.org/docs/trunk/components/rules/`

## References


1. Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods. *CoRR*, abs/0912.3747, 2009.
2. M. Banko and Oren Etzioni. The tradeoffs between open and traditional relation extraction. In *Annual Meeting of the ACL*, 2008.
3. M.W. Berry and M. Castellanos. *Survey of Text Mining II: Clustering, Classification and Retrieval*. Springer-Verlag, 2008.
4. I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (ACM-TKDD)*, 2007.
5. Eva Blomqvist. Ontocase-automatic ontology enrichment based on ontology design patterns. In *International Semantic Web Conference*, pages 65–80, 2009.
6. Johan Bos. Wide-Coverage Semantic Analysis with Boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing*, pages 277–286. College Publications, 2008.
7. Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 25–32, New York, NY, USA, 2004. ACM.
8. M. Ciaramita and Y. Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP-06, Sydney, Australia*, 2006.
9. M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, 2005.
10. Philipp Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, 2006.
11. Philipp Cimiano and Johanna Vlker. Text2onto - a framework for ontology learning and data-driven change discovery, 2005.
12. Bonaventura Coppola, Aldo Gangemi, Alfio Massimiliano Gliozzo, Davide Picca, and Valentina Presutti. Frame detection over the semantic web. In Lora Aroyo et al., editor, *ESWC*, volume 5554 of *LNCS*, pages 126–142. Springer, 2009.
13. Jesse Davis and Pedro Domingos. Deep transfer: A markov logic approach. *AI Magazine*, 32(1):51–53, 2011.
14. Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proc. of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK, 2011.
15. G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. Lexical markup framework (LMF). In *Proc. of the International Conference on Language Resources and Evaluation (LREC) , Genoa, Italy. ACL*, 2006.
16. A. Gangemi and V. Presutti. Ontology Design Patterns. In S. Staab and R. Studer, editors, *Handbook on Ontologies, 2nd Edition*. Springer Verlag, 2009.
17. 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.
18. Silvana Hartmann, György Szarvas, and Iryna Gurevych. Mining multiword terms from wikipedia. In Maria Teresa Pazienza and Armando Stellato, editors, *Semi-Automatic Ontology Development: Processes and Resources*, pages 226–258. IGI Global, Hershey, PA, USA, 2012.

19. Sebastian Hellmann, Jens Lehmann, and Sören Auer. Linked-data aware uri schemes for referencing text fragments. In *EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer, 2012.
20. Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. An overview of event extraction from text. In *Proceedings of Derive2011 Workshop, Bonn*, 2011.
21. Hans Kamp. A theory of truth and semantic representation. In Jeroen A. G. Groenendijk, Teo M. V. Janssen, and Martin B. J. Stokhof, editors, *Formal Methods in the Study of Language*, volume 1, pages 277–322. Mathematisch Centrum, 1981.
22. Jens Lehmann, Chris Bizer, Georgi Kobilarov, Sren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3):154–165, 2009.
23. Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16:pp. 72–79, March-April 2001.
24. J. McCrae, D. Spohr, and P. Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference*, 2011.
25. A. Miles and S. Bechhofer. Skos simple knowledge organization system extension for labels (skos-xl). W3C Recommendation, http://www.w3.org/TR/skos-reference/skos-xl.html, 2009.
26. A. Moschitti, D. Pighin, and : . R. Basili (2008): . Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193224, 2008.
27. David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Journal of Linguisticae Investigationes*, 30:1, 2007.
28. Roberto Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2), 2009.
29. Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In *ACL*, 2010.
30. A. G. Nuzzolese, A. Gangemi, and V. Presutti. Gathering Lexical Linked Data and Knowledge Patterns from FrameNet. In *Proc. of the 6th International Conference on Knowledge Capture (K-CAP)*, pages 41–48, Banff, Alberta, Canada, 2011.
31. Maria Teresa Pazienza and Armando Stellato. *Semi-Automatic Ontology Development: Processes and Resources*. IGI Global, Hershey, PA, USA, 2012.
32. W. Peters, E. Montiel-Ponsoda, G. Aguado de Cea, and A. Gomez-Perez. Localizing ontologies in owl. In *Proceedings of OntoLex Workshop, http://olp.dfki.de/OntoLex07/*, 2007.
33. Simone Paolo Ponzetto and Michael Strube. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175, 2011.
34. Valentina Presutti, Francesco Draicchio, and Aldo Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In *EKAW: Knowledge Engineering and Knowledge Management that matters*. Springer, 2012.
35. Johanna Völker and Sebastian Rudolph. Lexico-logical acquisition of owl dl axioms – an integrated approach to ontology refinement, 2008.
36. Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. Aida: An online tool for accurate disambiguation of named entities in text and tables. In *Proceedings of the 37th International Conference on Very Large Databases, VLDB 2011*, Seattle, WA, US, 2011.