Chainable and Extendable Knowledge Integration Web Services

Felix Sasaki[1] and Milan Dojchinovski[2,3]

[1] DFKI
DFKI felix.sasaki@dfki.de
[2] Knowledge Integration and Language Technologies (KILT/AKSW)
InfAI, Leipzig University, Germany
dojchinovski@informatik.uni-leipzig.de
[3] Web Intelligence Research Group
FIT, Czech Technical University in Prague
firstname.lastname@fit.cvut.cz

Abstract. This paper introduces the current state of the FREME framework. We will put FREME into the context of linguistic linked data and related approaches of multilingual and semantic processing. In addition, we focus on two specific aspects of FREME: the FREME NER e-Service, and chaining of e-Services. We believe that the flexible and distributed combination of e-Services bears a potential for their mutual improvement.

Keywords: Linguistic Linked Data, NIF, NLP, named entity recognition, workflows, multilingual and semantic enrichment

1 Introduction

This paper presents the current state of FREME, a framework for multilingual and semantic enrichment of digital content. A detailed, general overview of the goals of FREME has been given in [9]. Here we will focus on two aspects of FREME: the FREME NER service and chaining of FREME services.

FREME is developed in the EU funded FREME project[4], which started in February 2015 and lasts for two years. The project has two aspects: the development of the FREME framework, transferring technology outcomes from several language and data related projects; and the following four business cases:

1. Authoring and publishing multilingually and semantically enriched eBooks;
2. Integrating semantic enrichment into multilingual content in translation and localisation;

[4] See http://www.freme-project.eu

3. Enhancing the cross-language sharing and access to open agricultural and food data; and
4. FREME-empowered personalised content recommendations.

The paper is structured as follows. In section 2, we set FREME into the context of the Keki workshop. In section 3, we provide a general overview of the FREME architecture. In section 4, we elaborate on the FREME NER service. In section 5, we discuss how this and other services can be chained together. In section 6, we conclude the paper.

2 FREME in Context

The development of the FREME framework can be described a) in the context of linguistic linked data, and b) with regards to challenges that arise from the four business cases.

Data related to linguistic and natural language processing. In the paradigm of linguistic linked data, more and more language resources are being published as part of the linguistic linked open data cloud[5]. FREME allows to process data available in the cloud as part of content enrichment workflows, for example to adapt named entity recognition with domain specific data sets.

Linguistic and NLP Ontologies. The LLOD cloud gatheres language resources that are represented with standard formats. FREME enrichment workflows make use of the following formats:

- The Natural Language Processing Interchange Format (NIF) [6] to represent data and enrichment information;
- The Internationalization Tag Set (ITS) 2.0[6] to represent metadata for improvement of enrichment workflows; and
- The OntoLex Lemon model [7] to represent lexica, including their meaning with respect to ontologies.

Linguistic linked open data workflows. The LLOD technology stack allows to create NLP and data services in a distributed and decentralized manner. FREME implements this stack by making use of the forehand described standards, and by adding a declarative approach to define and re-use enrichment workflows.

NLP techniques for knowledge extraction. One aim of LLOD is to provide techniques for knowledge extraction that deploy linked data. FREME

[5] See `http://linguistic-lod.org/llod-cloud` for a latest version of the LLOD cloud.

[6] See `https://www.w3.org/TR/its20/`

[7] See `https://www.w3.org/2016/05/ontolex/`

implements this approach in its FREME-NER service and allows users to adapt the service with custom datasets, again to be provided as linked data.

Approaches using mappings and their maintenance from semistructured sources. Industry applications of NLP and data enrichment workflows have to deal with a plethora of content formats. Semistructured formats like HTML or certain XML formats are widely used in applications. Via its e-Internationalization service, FREME allows to process these formats, not only for extraction, but for round-tripping, that is: storage of enrichment information in the original format.

The LLOD context of FREME can also be described from the point of view of the four business cases. From the business case perspective, several challenges arise when creating NLP and data processing applications. They are addressed in the following manner by FREME.

Interoperability and chainability. Applications often are provided as silo solutions. Integration of new functionality is then a time consuming task with high integration costs. By using the forehand described, standardized technology stack, this effort is reduced significantly. Details are described in section 5.

Adaptability. There is a growing set of applications for key NLP tasks like named entity recognition, see e.g. [7]. Many of them rely on the DBpedia dataset [1] for entity linking. Tools like Stanford NER [5] allow users to load their own dataset and prepare it for NER. However, for users without a technological background in NLP, it is very hard to adapt these tools. FREME eases the adaptation process in several ways: it eases ease of configuringthe configuration of enrichment workflows; and it eases ease using custom data sets and tailoring NER processing towards domains. Details for this adaptation are described in Section 4.

Data formats. The four business cases require enrichment workflows in many formats. For example, in localization the XML based XLIFF format [8] is widely used. Current multilingual and semantic applications allow extraction of content of such formats. However, for real-life applications, the enrichment information has to be stored inside the format, without breaking existing processing tasks like validation, query or transformation. Via the e-Internationalization service, FREME allows such round-tripping processing.

2.1 Related Work

In this section we compare FREME to two related approaches: Apache Stanbol and Big Data Europe. They offer related capabilities and a comparison helps to understand the role of FREME.

[8] See `http://docs.oasis-open.org/xliff/xliff-core/v2.0/xliff-core-v2.0.html`

Apache Stanbol offers a set of text analytics services as Software as a Service using an open source platform. It intends to extend traditional content management systems with semantic services. Further the text analytics services can be used independently from arbitrary applications [2].

Apache Stanbol differs from FREME with regards to the set of services being offered. Although being open source and therefore being theoretically extensible Apache Stanbol, offers no detailed documentation on how to extend it. Further, Apache Stanbol puts a focus on the use case of a semantic content management system and semantic enrichment of homepages. Multilingual enrichment of other types of content is not taken into account.

The *Big Data Europe*[9] (BDE) project aims to develop an adaptable big data platform that is easy to deploy and use, allowing interested stakeholders to extend their big data solutions or introduce big data technology into their business processes. One focus of the project lies within the development of a multi-purpose infrastructure that is not tied to a specific big data technology or to a specific domain. The goal is a "one size fits all" system that unifies many existing technologies. Therefore it works on a high level of abstraction. It solves problems like managing large clusters that run concurrent big data pipelines.

In contrast to the abstraction of BDE, FREME sets various concrete aspects. We focus on a specific data type (textual content), on specific processes in the area of semantic technologies (like named entity recognition) and multilingual technologies (like machine translation), and on the challenge to enrich selected content formats. Hence, it might be possible to make FREME a component of the Big Data Europe architecture. This idea needs further evaluation and is out of scope for this paper.

3 FREME Architecture

FREME uses a client-server Web service architecture that exposes Web services, called *e-Services*, via HTTP APIs. This approach allows for a decentralized, distributed creation of services in a RESTful architecture [4]. In this way a combination of services does not need to be hard-wired and is not bound to a specific programming language, since almost every programing language supports HTTP based interactions [8]. Additionally, it is designed in an extensible manner, so that both project partners as well as external partners are able to add more services.

FREME uses common formats for language and data processing workflows, so that e-Services can easily be created by following the linguistic linked data technology stack. In this stack, the Natural Language Interchange Format (NIF) serves as a common broker format. Both the actual

[9] https://www.big-data-europe.eu

textual content and information generated via NLP and Linked Data processes is stored in NIF.

FREME offers six e-Services. Their functionality is summarized below.

- e-Enity offers named entity recognition. It is discussed in detail in section 4.
- e-Translation offers cloud based machine translation.
- e-Terminology offers enrichment of content with information about terms.
- e-Link offers enrichment with information from the linked data cloud.
- e-Publishing allows to store enriched content in the standardised ePub format.
- e-Internationalisation allows enrichment covering a wide range of digital content formats like HTML, generic XML or selected XML vocabularies.

In addition, the FREME framework is deployed in the German project "Digitial Curation Technologies" (DKT) [10]. Services offered by DKT also use the linguistic linked data technology stack and hence can be combined with FREME services out of the box.

4 Content Enrichment with Names Entities

4.1 FREME NER Overview

E-entity is one of the most exploited service within the FREME framework. Knowing what entities are mentioned in a document is of essential importance to better understand the aboutness of the document. The e-entity service annotates an input document with annotations representing entities. Mentions of entities, such as people, organizations or locations, are *spotted* and encoded with their position in the input document. Next, the entity is *disambiguated* by classifying from a set of entity types[11] and linking it to a specified knowledge base. The spotting and classification step is done by employing the StanfordNER tool [12][5] with a trained models on content from Wikipedia. The linking of entities ultimately relies on the most-frequent-sense approach and links with the most-frequent-sense entity. FREME NER is currently using models trained for English, German, Dutch, Spanish, Italian, French and Russian. To realize a MFS based linking we used Wikipedia as a reference knowledge base and collected every entity surface form, the corresponding hyperlink and the number of occurrences. As a result, a pair-count dataset [3] which provides these information was

[10] See `http://digitale-kuratierung.de/` for details on the project

[11] Currently, FREME classifies the entities with four types: PER, ORG, LOC and MISC for anything else.

[12] `http://nlp.stanford.edu/software/CRF-NER.shtml`

generated. The linking step is implemented in Apache Solr[13]. In Solr are indexed entities with their corresponding URI identifier, possible surface form variations, language, and the dataset they refer to. When performing the linking step, for an entity mention entity candidates are retrieved according to their surface form similarity, and the one with the highest `pair count` value is considered as the correct entity.

The listing 1.1 provides an example of the output from FREME NER.

```
<http://freme-project.eu/#char=0,33>
        a               nif:String , nif:Context , nif:RFC5147String ;
        nif:beginIndex  "0"^^xsd:int ;
        nif:endIndex    "33"^^xsd:int ;
        nif:isString    "Diego Maradona is from Argentina."^^xsd:string .

<http://freme-project.eu/#char=0,14>
        a                     nif:Word , nif:String , nif:Phrase , nif:RFC5147String ;
        nif:anchorOf          "Diego Maradona"^^xsd:string ;
        nif:beginIndex        "0"^^xsd:int ;
        nif:endIndex          "14"^^xsd:int ;
        nif:referenceContext  <http://freme-project.eu/#char=0,33> ;
        itsrdf:taClassRef     <http://dbpedia.org/ontology/SportsManager> , <http://
     dbpedia.org/ontology/Person> ... ;
        itsrdf:taConfidence   "0.9869992701528016"^^xsd:double ;
        itsrdf:taIdentRef     <http://dbpedia.org/resource/Diego_Maradona> .

<http://freme-project.eu/#char=23,32>
        a                     nif:String , nif:Word , nif:Phrase , nif:RFC5147String ;
        nif:anchorOf          "Argentina"^^xsd:string ;
        nif:beginIndex        "23"^^xsd:int ;
        nif:endIndex          "32"^^xsd:int ;
        nif:referenceContext  <http://freme-project.eu/#char=0,33> ;
        itsrdf:taClassRef     <http://dbpedia.org/ontology/Place> , <http://dbpedia.org/
     ontology/Location> ... ;
        itsrdf:taConfidence   "0.9804963628413852"^^xsd:double ;
        itsrdf:taIdentRef     <http://dbpedia.org/resource/Argentina> .
```

Listing 1.1. Output from FREME NER in the NIF format.

4.2 Entity Linking with Custom Datasets

In the last decade, entity linking has been primarily evaluated on datasets such as DBpedia, YAGO[14] and BabelNet[15]. In these use cases, the entity linking approaches have been exclusively customized to these datasets, and adoption of other datasets requires significant amount of effort, or it is not possible at all. In FREME, we allow users to use their custom proprietary and public datasets and adopt the processing according to their needs.

FREME NER provides a dataset management endpoint which can be used to perform the usual dataset operations such as creation, update and deletion of a dataset. The mimimum requirement is to provide a list of entities with a corresponding name variations. These information should be

[13] `http://lucene.apache.org/solr/`

[14] `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/`

[15] `http://babelnet.org/`

provided in RDF, where the subject of a triple is a URI, which uniquely identifies the entity, and the object is the entity name variation. The name variations can be provided using the RDFS[16] property `rdfs:label` or the SKOS[17] properties `skos:prefLabel` or `skos:altLabel`.

4.3 Domain Specific NER

In long texts, the list of recognized entities can be very large containing also entities which are not relevant to the domain of the document. For example, very often an HTML content contains also advertisements in from of text snippets which occur inline with the main content, and contain entity mentions which are irrelevant for the main content. In an HTML document providing recent information about the Syrian crisis might occur an advertisement related to the UEFA Euro 2016 championship, for example, can be mentioned a football team or football player. FREME enables users to filter out such irrelevant entities by specifying the domain of interest (i.e. Sports) Thus, only entities from this specific domain will be returned. The implementation of this feature is realized by populating list of domains with corresponding entity types. E.g. the types `http://dbpedia.org/ontology/SportsClub` and `http://dbpedia.org/ontology/SportsTeam` for the Sports domain.

4.4 Experiments

In order to evaluate the 1) *quality* of the enrichments and the 2) *scalability* of the named entity recognition, we have conducted several experiments using the GERBIL[18][11] framework. The experiments were executed in a local environment using GERBIL version 1.2.0-SNAPSHOT. The entity recognition was evaluated on five English and one German collection. The collections differ in length of the documents, the density of entity mentions and the topic of the documents[19]. Two types of experiments were executed; strong annotation match–requires exact match of the entity mention with the gold standard; and weak annotation match–requires overlap of the entity mention with the annotation in the gold standard. Table 1 summarizes the results from the experiments.

The results show that quality of the enrichments depend on the content. The best performance were achieved for the DBpedia Spotlight dataset with `0.944` F1 for the weak annotation match and 0.883 F1 for strong annotation match. The dataset contains 60 natural language sentences from ten different New York Times articles. The worst performance have been achieved for

[16] `https://www.w3.org/TR/rdf-schema/`

[17] `https://www.w3.org/TR/skos-reference/`

[18] `http://aksw.org/Projects/GERBIL.html`

[19] Please refer to [10] for more information on the datasets.

Table 1. Evaluation results of FREME NER.

Dataset	**Lang.**	**Exp. type**	**micro F1**	**micro P**	**micro R**	**macro F1**	**macro P**	**macro R**	**Avg millis per doc**	**Avg entities per doc**
Spotlight	EN	weak	**0.944**	0.944	0.944	**0.937**	0.958	0.940	67.02	5.69
		strong	0.882	0.882	0.882	0.872	0.891	0.876	47.90	5.69
KORE50	EN	weak	**0.944**	0.944	0.944	**0.937**	0.958	0.940	45.78	2.86
		strong	0.882	0.882	0.882	0.872	0.891	0.876	45.74	2.86
Reuters-128	EN	weak	0.742	0.732	0.753	0.735	0.681	0.847	113.59	4.85
		strong	0.612	0.603	0.621	0.606	0.561	0.699	100.92	4.85
RSS-500	EN	weak	0.680	0.529	0.951	0.730	0.638	0.951	67.76	0.99
		strong	0.578	0.450	0.809	0.626	0.548	0.809	66.47	0.99
MSNBC	EN	weak	0.825	0.867	0.787	0.793	0.759	0.839	734.94	32.50
		strong	0.728	0.765	0.694	0.692	0.664	0.732	894.78	32.50
News-100	DE	weak	**0.644**	0.777	0.550	**0.587**	0.631	0.571	369.73	22.33
		strong	0.447	0.535	0.384	0.373	0.398	0.365	232.42	14.04

the Reuters-128 dataset, which contains economic news articles. According to our observations, the DBpedia Spotlight content is very similar to the content we used to train our NER models, which is the reason for such good results. Reuters-128 content is from the financial domain which might be the reason for the lower performance.

In the experiments we have also evaluated the scalability of the entity recognition, and the evaluation results show that FREME NER in average can process one entity in 28 milliseconds or, in other words, 35 entities per second. Note that this conclusions, should be take with some reserve, since we implement caching and documents with frequently occurring entities will be processed faster.

5 Chainable Web Services

As described previously, FREME NER is just one e-Service provided by FREME. A key benefit of FREME is its approach to combine e-Services. This will be explained with the example in listing 1.2.

A pipeline consists of one or more steps. Each steps can take various input formats. If no format is specified, the step assumes NIF. The first step in the example pipeline evokes FREME NER which has been described in the previous section. The second step uses the e-link service to gather information with a selected query template. The third step calles the e-Terminology e-Service to enrich the content with terminology related information. This needs a source and a target language, here English and Dutch. The last step calls the e-Translation service, with the same language pairs.

The example pipeline shows several benefits. First, one can compare the outcome of several e-Services. In the example named entity recognition and terminology annotation are used to enrich the same content. This combination has the potential to improve both services via data based comparisons.

Second, there is no need to hardwire the combination of services, as long as the services adhere to the linguistic linked data stack. This can be seen in line 10 of the example. The e-Link service is installed at a different server (with the domain api.freme-project.eu) than the other e-services. The combination of services does not need a hardwired integration.

Third, the pipeline and in this way the e-Services are agnostic to given input and output formats. Format coverage is realised with the e-Internationalisation service. Separating the actual services and the formats to be processed has the advantage that other services easily can be integrated and benefit from the growing set of formats being supported.

Fourth, the pipelining greatly allows for automization of repetivive processes and for making the content itself intelligent. For example, a client application could analyze the content with regards to the language of content and use this information for adapting the pipeline automatically.

```
1   {
2   "id": 55,
3   "description": "Example pipeline",
4   "serializedRequests": [
5       {
6       "endpoint": "http://api-dev.freme-project.eu/current/e-entity/freme-ner/documents",
7       "parameters": {"language": "en"}
8       },
9       {
10      "endpoint": "http://api.freme-project.eu/current/e-link/documents/",
11      "parameters": {"templateid": "3"}
12      },
13      {
14      "endpoint": "http://api-dev.freme-project.eu/current/e-terminology/tilde",
15      "parameters": {
16      "source-lang": "en", "target-lang": "nl" }
17      },
18      {
19      "endpoint": "http://api-dev.freme-project.eu/current/e-translation/tilde",
20      "parameters": {
21      "source-lang": "en", "target-lang": "nl" }
22      }
23      ] }
```

Listing 1.2. Pipeline combining several e-Services

6 Conclusion

This paper introduced the current state of the FREME framework with regards to two aspects: named entity recognition via the FREME NER e-Service, and chaining of e-Services. In addition, we have put FREME into the context of linguistic linked data and related approaches of multilingual and semantic processing.

The discussion on FREME NER showed some preliminary evaluation results. The pipeling of e-Services has a practical benefit (e.g. ease and automization of similiar language and data processing workflows), but also a research potential. We think that the combination of named entity recognition, terminology annotation and machine translation can lead to a data driven improvement of all of these technologies. This is a potential next step for FREME.

Acknowledgement. This work was supported via the FREME project funded from the EU Horizon 2020 research and innovation programme under grant agreement No 644771.

References

1. S. Auer, *et al.* DBpedia: A Nucleus for a Web of Open Data. In *The semantic web*, pp. 722–735. Springer, 2007.
2. R. Bachmann-Gmur. *Instant Apache Stanbol.* Packt Publishing Ltd, 2013.
3. M. BrÃijmmer, M. Dojchinovski, and S. Hellmann. Dbpedia abstracts: A large-scale, open, multilingual nlp training corpus. In N. C. C. Chair), *et al.*, (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* European Language Resources Association (ELRA), Paris, France, may 2016.
4. R. T. Fielding and R. N. Taylor. Principled design of the modern web architecture. *ACM Transactions on Internet Technology (TOIT)*, 2(2):115–150, 2002.
5. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 363–370. Association for Computational Linguistics, 2005.
6. S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating NLP using linked data. In *International Semantic Web Conference*, pp. 98–113. Springer, 2013.
7. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pp. 1–8. ACM, 2011.
8. C. Pautasso, O. Zimmermann, and F. Leymann. Restful web services vs. big'web services: making the right architectural decision. In *Proceedings of the 17th international conference on World Wide Web*, pp. 805–814. ACM, 2008.
9. F. Sasaki, *et al.* Introducing freme: Deploying linguistic linked data. In *Proceedings of the 4th Workshop on the Multilingual Semantic Web*. 2015.
10. R. Usbeck, M. Röder, and A.-C. N. Ngonga. Evaluating entity annotators using gerbil. In *European Semantic Web Conference*, pp. 159–164. Springer, 2015.
11. R. Usbeck, *et al.* GERBIL – general entity annotation benchmark framework. In *24th WWW conference*. 2015.