Named Entity Recognition using FOX

René Speck and Axel-Cyrille Ngonga Ngomo

AKSW, Department of Computer Science, University of Leipzig, Germany
{speck,ngonga}@informatik.uni-leipzig.de

Abstract. Unstructured data still makes up an important portion of the Web. One key task towards transforming this unstructured data into structured data is named entity recognition. We demo FOX, the Federated knOwledge eXtraction framework, a highly accurate open-source framework that implements RESTful web services for named entity recognition. Our framework achieves a higher F-measure than state-of-the-art named entity recognition frameworks by combining the results of several approaches through ensemble learning. Moreover, it disambiguates and links named entities against DBpedia by relying on the AGDISTIS framework. As a result, FOX provides users with accurately disambiguated and linked named entities in several RDF serialization formats. We demonstrate the different interfaces implemented by FOX within use cases pertaining to extracting entities from news texts.

1 Introduction

The Semantic Web vision requires the data on the Web to be represented in a machine-readable format. Given that a significant percentage of the data available on the Web is unstructured, tools for transforming text into RDF are of central importance. In this demo paper, we present FOX, the federated knowledge extraction framework.[1] It integrates state-of-the-art named entity recognition (NER) frameworks by using ensemble learning (EL). By these means, FOX can achieve up to 95.23% F-measure where the best of the current state-of-the-art system (Stanford NER) achieves 91.68% F-measure. In this paper, we aim to demonstrate several of the features of FOX, including the large number of input and output formats it supports, different bindings with which FOX can be integrated into Java and Python code and the easy extension model underlying the framework. Our framework is already being used in several systems, including SCMS [5], ConTEXT [3] and IR frameworks [8]. The approach underlying FOX is presented in [7], which will be presented at the same conference. All features presented herein will be part of the demonstration.

2 Demonstration

The goal of the demonstration will be to show the whole of the FOX workflow from the gathering and preprocessing of input data to the generation of RDF data. In addition, we will show how to configure and train FOX after it has been enhanced with a

[1] FOX online demo:http://fox-demo.aksw.org
FOX project page:http://fox.aksw.org.
Source code, evaluation data and evaluation results:http://github.com/AKSW/FOX.

novel NER tool or EL algorithm. Further, we will present FOX's feedback RESTful service to improve the training and test datasets. In the demonstration, we also go over the Python[2] and Java bindings[3] for an easy use of FOX's RESTful service within an application. At the end we will explain how to use the FOX Java interfaces to integrate future algorithms.

2.1 Workflow

The workflow underlying FOX consists of four main steps: (1) preprocessing of the unstructured input data, (2) recognizing the Named Entities (NE), (3) linking the NE to resources using AGDISTIS [9] and (4) converting the results to an RDF serialization format.

Preprocessing FOX allows users to use a URL, text with HTML tags or plain text as input data (see the top left part of Figure 1). The input can be carried out in a form (see the center of Figure 1) or via FOX's web service. In case of a URL, FOX sends a request to the given URL to receive the input data. Then, for all input formats, FOX removes HTML tags and detects sentences and tokens.

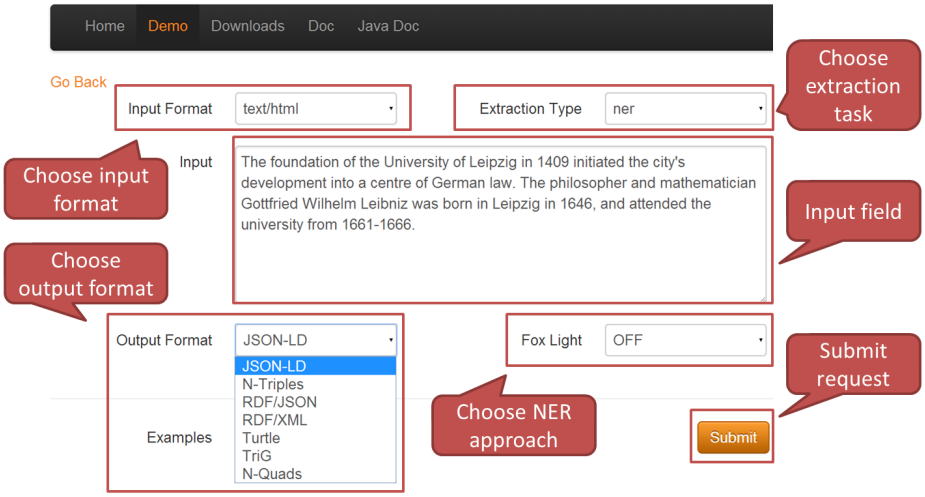

Fig. 1. Request form of the FOX online demo.

We will use text examples, URLs and text with HTML tags to show how FOX gathers or cleans them for the sake of entity recognition.

[2] `https://pypi.python.org/pypi/foxpy`
[3] `https://github.com/renespeck/fox-java`

Entity Recognition Our approach relies on four state-of-the-art NER tools so far: (1) the Stanford Named Entity Recognizer (Stanford) [2], (2) the Illinois Named Entity Tagger (Illinois) [6], (3) the Ottawa Baseline Information Extraction (Balie) [4] and (4) the Apache OpenNLP Name Finder (OpenNLP) [1]. FOX allows using a particular NER approach which is integrated in it (see bottom right of Figure 1). To this end, FOX light has to be set to the absolute path to the class of the tool to use. If FOX light is off, then FOX utilizes these four NER tools in parallel and stores the received NEs for further processing. It maps the entity types of each of the NER tools to the classes `Location`, `Organization` and `Person`. Finally, the results of all tools are merged by using FOX's EL layer as discussed in [7]. We will show the named entities recognized by FOX and contrast these with those recognized by the other tools. Moreover, we will show the runtime log that FOX generates to point to FOX's scalability.

Entity Linking FOX makes use of AGDISTIS [9], an open-source named entity disambiguation framework able to link entities against every linked data knowledge base, to disambiguate entities and to link them against DBpedia. In contrast to lookup-based approaches, our framework can also detect resources that are not in DBpedia. In this case, these are assigned their own URIs. Moreover, FOX provides a Java interface and a configuration file for easy integration of other entity linking tools. We will show the messages that FOX generates and sends to AGDISTIS as well as the answers it receives and serializes.

Serialization Formats FOX is designed to support a large number of use cases. To this end, our framework can serialize its results into the following formats: JSON-LD[4], N-Triples[5], RDF/JSON[6], RDF/XML[7], Turtle[8], TriG[9], N-Quads[10]. FOX allows the user to choose between these formats (see bottom left part of Figure 1). We will show how the out of FOX looks like in the different formats and point to how they can be parsed.

3 Evaluation and Results

We performed a thorough evaluation of FOX by using five different datasets and comparing it with state-of-the-art NER frameworks (see Table 1). Our evaluation shows that FOX clearly outperforms the state of the art. The details of the complete evaluation are presented in [7]. The evaluation code and datasets are also available at FOX's Github page, i.e., `http://github.com/AKSW/FOX`.

[4] `http://www.w3.org/TR/json-ld`
[5] `http://www.w3.org/TR/n-triples/`
[6] `http://www.w3.org/TR/rdf-json`
[7] `http://www.w3.org/TR/REC-rdf-syntax`
[8] `http://www.w3.org/TR/turtle`
[9] `http://www.w3.org/TR/trig`
[10] `http://www.w3.org/TR/n-quads`

Table 1. Comparison of the F-measure of FOX with the included NER tools. Best results are marked in bold font.

	token-based					entity-based				
	News	News*	Web	Reuters	All	News	News*	Web	Reuters	All
FOX	**92.73**	**95.23**	**68.81**	**87.55**	**90.99**	**90.70**	**93.09**	**63.36**	81.98	**90.28**
Stanford	90.34	91.68	65.81	82.85	89.21	87.66	89.72	62.83	79.68	88.05
Illinois	80.20	84.95	64.44	85.35	79.54	76.71	83.34	54.25	**83.74**	76.25
OpenNLP	73.71	79.57	49.18	73.96	72.65	67.89	75.78	43.99	72.89	67.66
Balie	71.54	79.80	40.15	64.78	69.40	69.66	80.48	35.07	68.71	67.82

4 Conclusion

We will present FOX, a NER framework which relies on EL and demonstrate how it can be used. In future work, we will extend the number of tools integrated in FOX. Moreover, we will extend the tasks supported by the framework. In particular, we aim to integrate tagging, keyword extraction as well as relation extraction in the near future.

References

1. J Baldridge. The opennlp project, 2005.
2. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370, 2005.
3. Ali Khalili, Sören Auer, and Axel-Cyrille Ngonga Ngomo. context – lightweight text analytics using linked data. In *11th Extended Semantic Web Conference (ESWC2014)*, 2014.
4. David Nadeau. Balie—baseline information extraction: Multilingual information extraction from text with machine learning and natural language techniques. Technical report, Technical report, University of Ottawa, 2005.
5. Axel-Cyrille Ngonga Ngomo, Norman Heino, Klaus Lyko, René Speck, and Martin Kaltenböck. SCMS - Semantifying Content Management Systems. In *Proceedings of the International Semantic Web Conference*, 2011.
6. Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
7. René Speck and Axel-Cyrille Ngonga Ngomo. Ensemble learning for named entity recognition. In *In Proceedings of the International Semantic Web Conference*, Lecture Notes in Computer Science, 2014.
8. Ricardo Usbeck. Combining linked data and statistical information retrieval. In *11th Extended Semantic Web Conference, PhD Symposium*. Springer, 2014.
9. Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Sören Auer, Daniel Gerber, and Andreas Both. Agdistis - agnostic disambiguation of named entities using linked open data. In *Proceedings of 13th International Semantic Web Conference*, 2014.