Title of the LREC 2012 Paper

Author1, Author2, Author3

Affiliation1, Affiliation2, Affiliation3
Address1, Address2, Address3
author1@xxx.yy, author2@zzz.edu, author3@hhh.com

Abstract

the main idea is to give an overview of the available benchmarks:
the story is:
- recently uptake of NIF for benchmarks
- general description from page 84 of http://svn.aksw.org/papers/2013/Thesis_Sebastian/
- overview of existing benchmarks:
- validation with databugger

Keywords: NLP, Linked Data, Benchmark, Validation

1. Introduction

2. Background(sebastians thesis)

2.1. NIF

2.2. NER Extension of NIF

2.3. Linked Data Principles in NIF (Lim)

vide a
mmary
the pa-
s and
data

3. Existing corpora

3.1. N3 (Ricardo)

http://aksw.org/Projects/N3NERNEDNIF.html

3.2. Magnus

(Steinmetz et al., 2013): DBpedia Spotlight dataset, KORE 50 (AIDA), Wikilinks Corpus (Singh et al., 2012) subset (triplified by AKSW)

3.3. Wikilinks(Martin)

The NIF conversion of the Wikilinks corpus, as described in (Hellmann et al., 2013), could be improved by using the expanded dataset. Now every item of the corpus contains the full DOM of a website, its URI as well as a number of mentions that link to the English Wikipedia, including the link text and the context string. It still is very large in scale, containing over 3 million items and 40 million mentions. However, the compressed size has grown to 180GB, making it much harder to handle, but at the same time granting much better conversion opportunities. For instance, the complete DOM structures can be used to identify the context of the mentions much better and extract more text useable for NER disambiguation. You can see an example in Listing 1

The new NIF conversion establishes one `nif:Context` per item, instead of one per mention, like before. The DOM of the website is parsed and every mention's link is found to extract the relevant surrounding HTML element's text. This results in a clean and semantically relevant text snippet for each mention, instead of the arbitrary context strings of fixed length that where used before. In addition to linking DBpedia via `itsrdf:taIdentRef`, DBpedia ontology types[1] where included for every mention having a DBpedia ontology type via `itsrdf:taClassRef`. NERD classes directly mapping the DBpedia ontology types where also included via `itsrdf:taClassRef`. To be able to directly identify a coarse grained instance type (i.e.Person, Location, Organization, etc.), the NERD core class containing the mapped NERD class was added via `nif:taNerdCoreClassRef`.

```
<http://wiki-link.nlp2rdf.org/linkeddata.php?t=url&f=
  html&i=http://www.methodinit.org.uk/methodinit/2007/
  11#char=0,8353>
  a nif:String , nif:Context , nif:RFC5147String ;
  nif:isString """A Libertarian and Relativist quote
    taken from the Christian Anarchist Leo Tolstoy .
    Somewhat compatible with discourse
    theory."""^^xsd:string;
  nif:beginIndex "0"^^xsd:nonNegativeInteger;
  nif:endIndex "8353"^^xsd:nonNegativeInteger;
  nif:sourceUrl
    <http://www.methodinit.org.uk/methodinit/2007/11> .

<http://wiki-link.nlp2rdf.org/linkeddata.php?t=url&f=
  html&i=http://www.methodinit.org.uk/methodinit/2007/
  11#char=70,81>
  a nif:String , nif:RFC5147String ;
  nif:referenceContext <http://wiki-link.nlp2rdf.org/
    linkeddata.php?t=url&f=html&i=
    http://www.methodinit.org.uk/methodinit/2007/11
    #char=0,8353> ;
  nif:anchorOf """Leo Tolstoy"""^^xsd:string ;
  nif:beginIndex "70"^^xsd:nonNegativeInteger ;
  nif:endIndex "81"^^xsd:nonNegativeInteger ;
  a nif:Phrase ;
  itsrdf:taClassRef
    <http://dbpedia.org/ontology/Writer> ;
  itsrdf:taClassRef
    <http://dbpedia.org/ontology/Artist> ;
  itsrdf:taClassRef
    <http://nerd.eurecom.fr/ontology#Artist> ;
  itsrdf:taClassRef
    <http://dbpedia.org/ontology/Person> ;
  itsrdf:taClassRef
    <http://dbpedia.org/ontology/Agent> ;
  nif:taNerdCoreClassRef
    <http://nerd.eurecom.fr/ontology#Person> ;
  itsrdf:taIdentRef
```

[1] http://downloads.dbpedia.org/3.9/en/instance_types_en.nt.bz2

```
    <http://dbpedia.org/resource/Leo_Tolstoy> .
```

Listing 1: A converted wikilinks item including one mention

3.4. wikipedia corpus (Lim with help from Felix, Dimitris)

- our wikipedia corpus, i.e. felix xslt script (= Wikilinks Corpus ?)

3.5. Overview + Table (Lim)

`http://svn.aksw.org/papers/2014/ESWC_NLP_Cleansing/`

4. Validation (Dimitris)

5. Towards Standardized NER Benchmarking based on Gate (Milan)

6. Related Work and Conclusions

7. Acknowledgements

Place all acknowledgements (including those concerning research grants and funding) in a separate section at the end of the article.

8. References

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer, 2013. Integrating nlp using linked data. In *12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*.

Singh, Sameer, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum, 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015, University of Massachusetts, Amherst.

Steinmetz, Nadine, Magnus Knuth, and Harald Sack, 2013. Statistical analyses of named entity disambiguation benchmarks. In *Proceedings of 1st International Workshop on NLP and DBpedia*, volume 1064 of *CEUR Workshop Proceedings*. Sydney, Australia: CEUR-WS.org.