HeiNER

Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration

Wolodja Wentland, Johannes Knopp, Carina Silberer and Matthias Hartung
Department of Computational Linguistics, University of Heidelberg

1. Introduction

HeiNER - **Hei**delberg **N**amed **E**ntity **R**esource

Multilingual resource for Named Entity Disambiguation, Translation and Transliteration, which is freely available at http://heiner.cl.uni-heidelberg.de

Contains

- 1.547.586 disambiguated English NEs
- Translations into 253 languages
- Context Sets in 16 languages
- A dictionary that maps ambiguous proper names to sets of unique and disambiguated NEs

The resource is automatically constructed from Wikipedia by exploiting its internal link structure.

Method

- Based on Wikipedia
- Extraction of disambiguated NEs
- Translation of monolingual NE seeds into all languages available in Wikipedia
- Context Extraction in all languages

Advantages

- Avoids manual annotation
- Viable for resource-poor languages
- Yields large NE resource for multiple languages

Current State of the project

- Capitalisation-based heuristic for NER in English as source language, with a precision of **0.95**
- NE translation to 15 target languages

Usage

HeiNER provides data for supervised training in:

- Multilingual Information Extraction
- Machine Translation
- NER and NE Disambiguation
- Text summarisation

2. Named Entity Acquisition

Monolingual

Only article titles are considered as candidates, which circumvents the need for NE boundary detection.

Heuristic:

- Based on (Bunescu and Pasca, 2006)
- Three steps in recognition:

1. Multiword title “United_Nations”
2. CamelCase “YouTube”
3. Capitalisation in text

Capitalisation:

- Candidate is frequently capitalised (> 75%)
- Occurrences in sentence initial position are not counted

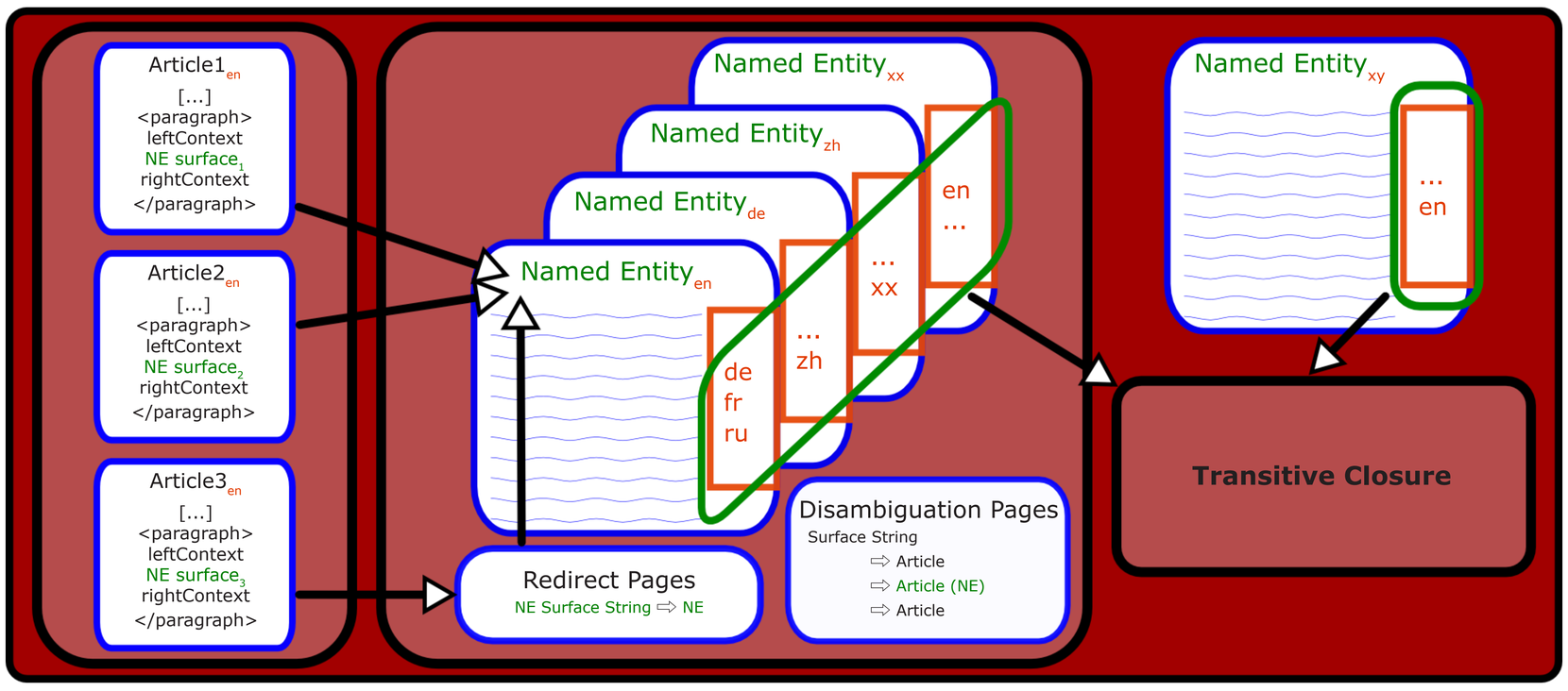
Disambiguation:

- Extraction of different - disambiguated - surface forms, by exploitation of Disambiguation and Redirect Pages in Wikipedia
- Disambiguation Pages distinguish between different NE readings and Redirect Pages unify different surface forms in a single representation
- Mappings between surface and disambiguated forms are stored in the **Disambiguation Dictionary**

Multilingual

Translation/Transliteration

- Utilisation of cross-language links in Wikipedia
- Translation of NEs from a single source language into all other languages in Wikipedia
- **Triangulation** between languages to increase coverage, by exploiting the fact that links should obey the principle of transitivity
- All extracted translations constitute the **Translation Dictionary**




```
<dataset>
 <context id=”0” article=”WHO”>
 <ne>United Nations</ne>
 <surfaceString>United Nations</surfaceString>
 <leftContext>
   The World Health Organization
  (WHO) is a specialized agency of the
 </leftContext>
 <rightContext>
   (UN) that acts as a coordinating
   authority on international public health.
 </rightContext>
 </context>
 ...
</dataset>
```

Context Set

```
<dictSet>
    <pageTitle>Python</pageTitle>
    <ne>Python (roller coaster)</ne>
    <ne>Colt Python</ne>
    <ne>Python II</ne>
    <ne>Armstrong Siddeley Python</ne>
    <ne>Python (film)</ne>
    <ne>Monty Python</ne>
    <ne>Python Automobile</ne>
    <ne>Python (programming language)</ne>
    <ne>Python (mythology)</ne>
    <ne>Pythonidae</ne>
</dictSet>
...
```

Disambiguation Dictionary

```
<transDict>
 <namedEntity id=’2134’>
  <an>Organizazión d’as Nazions Unitas</an>
  <bs>Ujedinjeni narodi</bs>
  <el>Οργανισμός Ηνωμένων Εθνών</el>
  <ga>Náisiúin Aontaithe</ga>
  <gl>ONU</gl>
  <he>ועא</he>
  <hu>Egyesült Nemzetek Szervezete</hu>
  <lb>Vereent Natiounen</lb>
  <nds>Vereente Natschonen</nds>
  <ru>Организация Объединённых Наций</ru>
  <tr>Birleşmiş Milletler</tr>
  <en>United Nations</en>
  ...
  <kr>국제 연합</kr>
  <ja>国際連合</ja>
  <zh>联合国</zh>
 </namedEntity>
  ...
</transDict>
```

Translation Dictionary

3. Context Acquisition

Contexts:

- Entire paragraphs are extracted
- Created for all target languages
- Contexts are extracted for disambiguated NEs, by means of the link structure

leftContext **[surfaceForm | NE]** *rightContext*

- All extracted contexts constitute the **Context Set** for a single language

Recognition:

- Search for links to NEs within all articles
- Ambiguous surface forms are disambiguated by means of the link structure

	NUMBER OF CONTEXTS			
Languages	ABSOLUTE	AMBIG. NE	MEAN	MEDIAN
de	9,665,648	1,573,173	50	6
en	43,065,047	8,076,626	51	5
es	4,198,613	499,051	76	6
fi	1,364,604	59,550	28	4
fr	7,627,032	1,096,986	60	6
it	4,819,325	517,419	54	5
ja	6,831,990	545,398	59	6
nl	3,784,999	590,949	25	6
no	1,345,096	86,519	28	2
pl	3,923,401	241,336	33	4
pt	2,853,306	365,945	53	7
ru	2,131,456	157,726	33	4
sv	1,898,103	172,796	55	4
sw	30,250	225	6	5
tr	625,072	21,468	37	4
zh	1,924,618	38,323	31	4
MEAN	**6,005,535**	**877,718**	**42**	**-**

Table 1: Number of contexts extracted for different languages

4. Results and Evaluation

Evaluation:

- 2 evaluation sets with 2000 Markables with two and three annotators respectively
- Fleiss Kappa Agreement of **0.774** and **0.771**
- Average NER Precision of **0.95**

Triangulation:

- Average boost in coverage of **4.47 %**
- Maximum boost in turkish with **7.73 %**
- **77,694** additional NE translations acquired

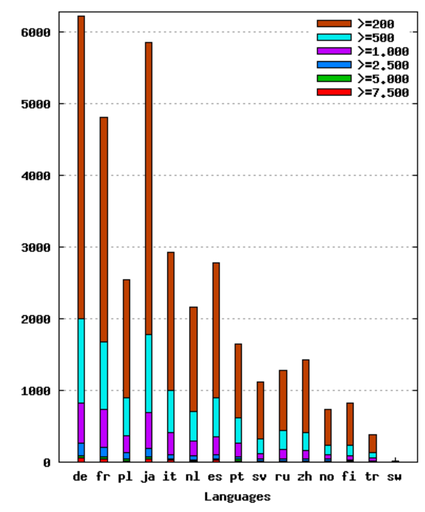
			DIFFERENCE	
Languages	INITIAL	FINAL	ABSOLUTE	PERCENT
de	243,903	250,049	6,146	2.46 %
es	127,518	137,606	10,088	7.33 %
fi	67,095	71,052	3,957	5.57 %
fr	215,479	222,712	7,233	3.25 %
it	135,852	145,889	10,037	6.88 %
ja	116,488	120,056	3,568	2.97 %
nl	166,708	176,203	9,495	5.39 %
no	63,431	66,786	3,355	5.02 %
pl	128,078	134,250	6,172	4.60 %
pt	132,778	137,227	4,449	3.24 %
ru	81,331	87,224	5,893	6.76 %
sv	97,270	99,710	2,440	2.45 %
sw	2,765	2,962	197	6.65 %
tr	26,814	29,059	2,245	7.73 %
zh	56,652	59,071	2,419	4.10 %
TOTAL	**1,662,162**	**1,739,856**	**77,694**	**4.47 %**

Table 2: Increase in coverage by means of triangulation

Context Distribution

- Contexts are unevenly distributed
- 1.5% of NEs in English cover 44% of the contexts
- ~23k NEs with more than 200 contexts
- Four NEs with 100,000+ contexts:

1. United States
2. England
3. United Kingdom
4. Germany



5. Further Work

- Providing the data as database dumps instead of XML
- Semantic Class labels for all NEs
- NER with proper NER systems in several languages
- Topic signatures for all NEs
- Context alignment