Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration

Wolodja Wentland **Johannes Knopp** **Carina Silberer** **Matthias Hartung**

Department of Computational Linguistics
Heidelberg University
{wentland, knopp, silberer, hartung}@cl.uni-heidelberg.de

Abstract

In this paper, we present *HeiNER*, the multilingual Heidelberg Named Entity Resource. HeiNER contains 1,547,586 disambiguated English Named Entities together with translations and transliterations to 15 languages. Our work builds on the approach described in (Bunescu and Pasca, 2006), yet extends it to a multilingual dimension. Translating Named Entities into the various target languages is carried out by exploiting crosslingual information contained in the online encyclopedia Wikipedia. In addition, HeiNER provides linguistic contexts for every NE in all target languages which makes it a valuable resource for multilingual Named Entity Recognition, Disambiguation and Classification. The results of our evaluation against the assessments of human annotators yield a high precision of 0.95 for the NEs we extract from the English Wikipedia. These source language NEs are thus very reliable seeds for our multilingual NE translation method.

1. Introduction

Named entities (NEs) are fundamental constituents in texts, but are usually unknown words. Thus, named entity recognition (NER) and disambiguation (NED) are a crucial prerequisite for successful information extraction, information retrieval, question answering, discourse analysis and machine translation. In line with (Bunescu, 2007), we consider NED as the task of identifying the unique entity that corresponds to potentially ambiguous occurrences of proper names in texts.[1] This is particularly important when information about a NE is supposed to be collected and integrated from multiple documents. Especially multilingual settings in information retrieval (Virga and Khudanpur, 2003; Gao et al., 2004) or question answering (Ligozat et al., 2006) require methods for transliteration and translation of NEs.

In this paper we propose a method to acquire NEs in one source language and translate and transliterate[2] them to a great variety of target languages. The method we apply in order to recognise and disambiguate NEs in the source language follow the approach of (Bunescu and Pasca, 2006). For multilingual NE acquisition we exploit and enrich crosslingual information contained in the online encyclopedia Wikipedia. With this account we demonstrate a viable solution for the efficient and reliable acquisition of disambiguated NEs that is particularly effective for resource-poor target languages, employing heuristic NER in a single source language. For the NED task, we rely on Wikipedia's internal link structures.

In its current state, HeiNER comprises the following components which are freely available[3]

- a *Translation Dictionary* of all NEs encountered in the English version of Wikipedia with their translations into 15 other languages available in Wikipedia,
- a *Disambiguation Dictionary* for each of the 16 languages, that maps all ambiguous proper names to the set of unique NEs they refer to, and
- a *Multilingual Context Database* for all disambiguated NEs.

From our perspective, it is especially the collection of linguistic contexts that are provided for the disambiguated NEs in every target language that makes HeiNER a valuable resource for NE-related tasks in multilingual settings. These contexts can be used for *supervised* training of classifiers for tasks such as NER, NED or NE Classification in languages for which no suitable systems are available so far.

The paper is structured as follows: In Section 2., we introduce the aspects of Wikipedia's internal structure that are essential for our approach. Section 3. describes the details of how we acquire NEs monolingually for one source language and translate them to many target languages. In order to assess the quality of HeiNER, we evaluate the seeds of the translation step, i.e. the set of NEs acquired in the source language, against the judgements of human annotators. The results of the evaluation are presented in Section 4. Section 5. concludes and gives an outlook on future work.

2. Wikipedia

2.1. Overview

Wikipedia[4] is an international project which is devoted to the creation of a multilingual free online encyclopedia. It uses freely available Wiki software[5] to allow users to create and edit its content collaboratively. This process relies

[1] As an example, (Bunescu, 2007) mentions the ambiguous proper name *Michael Jordan* which can either refer to the Basketball player or a University Professor at Berkeley.

[2] Note that throughout this paper we will use the term *translate* in a wider sense that subsumes *transliteration* into different scripts as well.

[3] http://heiner.cl.uni-heidelberg.de

[4] http://wikipedia.org

[5] http://www.mediawiki.org

entirely on subsequent edits by multiple users who are advised to adhere to Wikipedia's quality standards as stated in the *Manual of Style*[6].

Wikipedia has received a great deal of attention within the research community over the last years and has been successfully utilised in such diverse fields as machine translation (Alegria et al., 2006), NE transliteration (Sproat et al., 2006), word sense disambiguation (Mihalcea, 2007), parallel corpus construction (Adafre and de Rijke, 2006) and ontology construction (Ponzetto and Strube, 2007).

From a NLP perspective the attractiveness of employing the linguistic data provided by Wikipedia lies in the huge amount of NEs that it contains in contrast to commonly used lexical resources such as WordNet[7].

Wikipedia has grown impressively large over the last years. As of December 2007, Wikipedia had approximately 9.25 million articles in 253 languages, comprising a total of over 1.74 billion words for all Wikipedias[8]. Wikipedia versions for almost all major European languages, Japanese and Chinese have surpassed a volume of 100.000 articles (see Figures 1 and 2 for an overview).

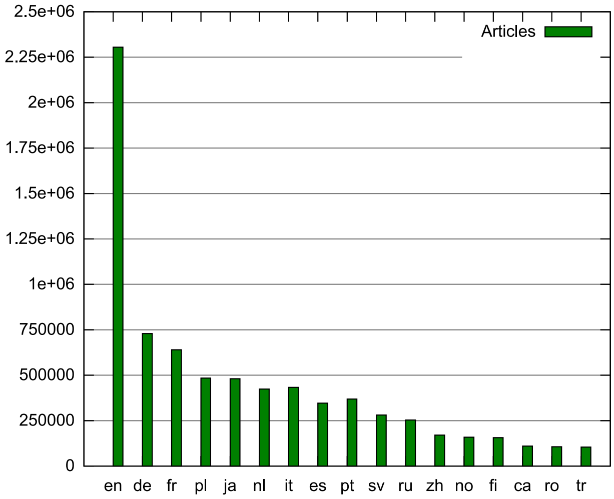

Figure 1: Number of articles found in different languages.

The enormous size of this comparable corpus[9] raised the question whether Wikipedia can be used to conduct multilingual research. We believe that the internal structure of Wikipedia and its multilingualism can be harnessed for several NLP tasks. In the following, we describe the technical properties of Wikipedia we make use of for the acquisition of multilingual NE information.[10]

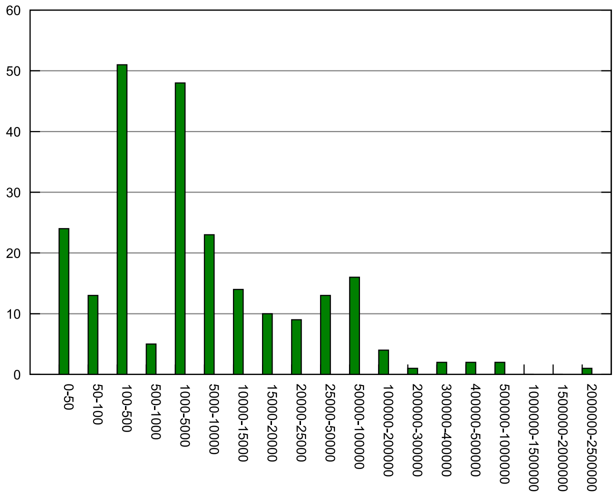

Figure 2: Language editions found in Wikipedia by number of articles.

2.2. Structure

Wikipedia is a large hypertext document which consists of individual articles, or pages, which are interconnected by links. The majority of links found in the articles is conceptual, rather than merely navigational. They often refer to more detailed descriptions of the respective entity[11] denoted by the link's anchor text (Adafre and de Rijke, 2005).

An important characteristic of articles found in Wikipedia is that they are usually restricted to the description of one single, unambiguous entity.

Pages

Wikipedia pages can be loosely divided into four classes: *Concept Pages*, *Named Entity Pages* and *Category Pages* account for most of Wikipedia's content and are joined by *Meta Pages* which are used for administrative tasks, inclusion of different media and disambiguation of ambiguous strings.[12]

Redirect Pages

Redirect Pages are used to normalise different surface forms, denominations and morphological variants of a given unambiguous NE or concept to a unique form or ID. To exemplify the use of Redirect Pages consider that `USA`, `United States of America` and `Yankee land` are all Redirect Pages to the article `United States`. These redirects enable us to identify `USA`, `United States of America` or `Yankee land` as varying surface forms of the unique NE `United States`.

[6] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`

[7] `http://wordnet.princeton.edu/`

[8] Data taken from `http://en.wikipedia.org/wiki/Wikipedia`

[9] As opposed to parallel corpora, comparable corpora are not strictly aligned, but contain topically related sentences. (Sproat et al., 2006)

[10] In doing so, we restrict ourselves to those properties we actually exploit for the purposes of this paper.

[11] Throughout this paper, we use this term as a generalization over all kinds of referents that are designated by Wikipedia pages, i.e. NEs, concepts, events, etc.

[12] The distinction between Named Entiy and Concept Pages is not reflected by Wikipedia's structure. Instead, it has to be accounted for by specific methods (see Section 3.1.1.).

Disambiguation Pages

Disambiguation Pages are used to disambiguate proper names to a set of distinct and unique NEs that may be denoted by them. The Disambiguation Page for `Python` lists articles discussing `Python (programming language)`, `Monty Python` and `Pythonidae`[13].

If a Disambiguation Page solely contains person names, geographical locations, schools, hospitals, places of worship, names of ships and roads or mathematical titles, users are advised to mark them as such. This enables us to extract additional lists[14] of classified NEs, as marked by these specific Disambiguation Pages.

Links

Just like links found on web pages, links in Wikipedia place a reference to another article within a text and can specify the surface form that will be displayed in the text. A typical link might look like `[[United Nations|UN]]`, where `UN` is the link's surface form while `United Nations` is the referenced article and link target.

Cross-language links

Cross-language links are links from any page describing an entity in one Wikipedia language to a page describing the same entity in another language. The cross-language links take the form `[[language code:Title]]`, where the language code is the two/three-letter code as per ISO 639.

We will exploit the existence of *cross-language links* within Wikipedia to acquire translations of NE readings from one to all other languages in Wikipedia.

2.3. Access to the Wikipedia data

For our task of identifying Named Entity pages and extracting alternative surface forms for these from all Redirect and Disambiguation pages that belong to a given NE, we have written a comprehensive Python API to access Wikipedia as a lexical semantic resource[15] which is also freely available.[16] Although our API was designed specifically for this project, its class library is easily extensible to account for different tasks and differentiates between a plethora of different page, link and template types in Wikipedia. Its design is almost completely language-independent[17], thus fostering the design of algorithms suited for multilingual processing.

3. Multilingual Resource Construction for HeiNER

3.1. Multilingual NE Acquisition

Our aim is to construct a multilingual NE resource comprising a disambiguation dictionary and a context dataset for every language available in Wikipedia.

As a necessary prerequisite, we need an initial list of NEs for every target language. There are two approaches to generate such lists:

- Implement a named entity recogniser or elaborate a language specific heuristic for NER for *every* target language.
- Develop a NER system or heuristic for NER for *one* source language and translate the NEs it recognises to the target languages of interest.

Due to language specific differences, NE recognisers are cumbersome to implement. Further, the performance of NER heuristics is often difficult to be assessed in advance. Therefore, we decided to follow the second approach for acquiring the multilingual components of HeiNER.

In the following subsections we describe the monolingual NER heuristic that we apply to the source language and how the resulting list of NEs is translated to the target languages by exploiting and enriching Wikipedia's crosslingual link structure.

3.1.1. Heuristic for Monolingual NE Acquisition in the Source Language

We chose English as the source language, for mainly two reasons: First, the English Wikipedia contains by far the largest amount of articles (cf. Figure 1). English therefore provides an optimal seed for multilingual NE acquisition in terms of recall. Second, linguistic insight reveals that the English language bears peculiarities (mainly capitalization patterns) which can be exploited as a considerably precise heuristic for NE recognition (see section 4.). Insofar, using English as our source language seems to be a good choice from the perspective of both recall and precision.

Note that our multilingual approach is (i) independent of the choice of a specific source language[18] and (ii) that any NER method can be applied for the inital NE acquisition in the source language.

As most of the Wikipedia articles refer to entities, there is per se a high probability for an article to describe a NE. Thus, by considering every article title[19] as a possible NE we avoid some of the non-trivial problems involved in NER, such as NE boundary detection (Nadeau et al., 2006). Therefore, it is not necessary to apply a sophisticated NER

[13] Family of non-venomous snakes

[14] These lists will be provided independently from the other resources and can be used as seed lists for other NER approaches.

[15] For other Wikipedia APIs for similar purposes, see (Strube and Ponzetto, 2006; Zesch et al., 2007).

[16] The source code and the documentation of the API can be downloaded from `http://heiner.cl.uni-heidelberg.de`.

[17] The only adaptations needed to account for other languages and scripts are (i) word segmentation in especially East Asian scripts and (ii) translation of Wikipedia namespaces (e.g. `Image:` is translated to `Bild:` in the German Wikipedia).

[18] In principle, any language other than English might be chosen (varying performance being the consequence, of course). In particular, it might turn out that a different source language performs even better for specific language pairs.

[19] The titles are normalised by removing additional information which is often contained in Wikipedia article titles. For instance, `The Firm (1993 film)` is normalised to `The Firm`.

system. Instead, a heuristic is applied on each article to determine if it corresponds to a NE or not. Additionally, by restricting our acquisition method to article titles, we avoid any efforts for morphological normalisation.

The heuristics we apply in order to acquire NEs in the source language are those suggested by (Bunescu and Pasca, 2006):
Let $e \in E$ be a given Wikipedia entity for which we have to decide whether it is an NE or not, and let *e.title* be the title attributed to this entity in Wikipedia:

1. If *e.title* is a multiword title, check the capitalization of all content words in *e.title*, i.e. words other than prepositions, determiners, conjunctions, relative pronouns or negations. Consider e a named entity if and only if all content words in the title are capitalized.

2. If *e.title* is a one word title that contains at least two capital letters, then *e* is a named entity. Otherwise, go to step 3.

3. Count how frequently *e.title* occurs in the text of the article that is related to e in positions other than at the beginning of sentences. If at least 75% of these occurrences are capitalized, then *e* is a named entity.

3.1.2. NE Translation from the Source Language to the Target Languages

After having identified the NEs occurring in the source language, we acquire translations of them into other languages by applying the following method: For all NEs extracted in the source language, mappings to corresponding NEs, i.e. their translations into the various target languages, are computed by exploiting Wikipedia's cross-language links (cf. section 2.2.). We represent these references in a translation/transliteration dictionary *transDict*: For each NE recognised in the source language, the corresponding translations/transliterations into the target languages are stored in the *transDict*. This method avoids the drawbacks of the alternative monolingual approaches sketched above.

3.1.3. Filling Linkage Holes by Triangulation

As our approach crucially depends on the density of Wikipedia's cross-language links, we implemented a method to fill linkage holes in order to increase the coverage of the acquisition step. The details of this method are described in the following.

A linkage hole means that there are corresponding NEs within two languages, but there is no direct link between the Wikipedia articles describing them. An example of this is illustrated in Figure 3. The figure shows three lists of cross-language links contained in the English (top left), French (down left) and Italian (right) article about 'Le Crestet'. Note that no link exists between the English and Italian Wikipedia pages. However, there is a link from the English to the corresponding French page and from the French to the corresponding Italian page. As a consequence, we add the Italian translation of 'Le Crestet' to our *transDict*. For the rest of the links the same inferencing can be applied, yet isn't illustrated here.

In order to achieve optimal density within Wikipedia's network of cross-language links, we make use of *triangulation*, a technique which is well-known from machine translation (Kay, 1997; Eisele, 2005). Generally, triangulation allows to infer a relation between lexical items of two different languages by detour via a third language. We apply triangulation to our problem of linkage holes by inferring missing cross-language links between languages L_1 and L_2, $l(ne_{L_1}, ne_{L_2})$, according to the following rule:

$$l(ne_{L_1}, ne_{L_3}) \wedge l(ne_{L_3}, ne_{L_2}) \Rightarrow l(ne_{L_1}, ne_{L_2})$$

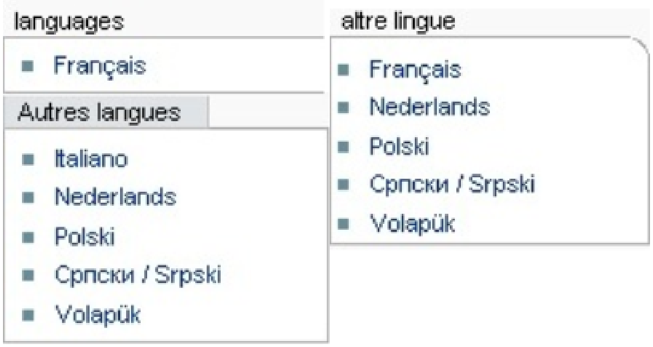

Figure 3: Linkage holes in cross-language links

The underlying assumption is that the network of cross-language links should, in fact, obey the principle of transitivity. Hence, in our approach, filling linkage holes by triangulation is accomplished by computing the transitive closure on the set of existing cross-language links within Wikipedia, thus collecting all NE translations in the *transDict*.

3.2. Creating the Disambiguation Dictionary and Context Dataset for Each Language

After extension by means of triangulation, the *transDict* serves as the base for building the final multilingual NE resource, comprising a disambiguation dictionary and a context dataset for every language available in Wikipedia. The methods for the creation follow in substance the settings from (Bunescu and Pasca, 2006). These are described below.

3.2.1. Creating the Disambiguation Dictionary

For each language, the *Disambiguation Dictionary* (*DisDict*) is constructed by executing the following steps:

1. for each NE page ne_i: update *DisDict* with an entry that contains ne_i*.title* as key and also as its value[20].

2. for all Disambiguation Pages d_i which contain one or more links to a NE ne_k: add the value ne_k to the *DisDict* entry with the key d_i*.title*[21].

[20] This mapping is inserted, as each NE can be referred to by itself, of course.

[21] The addition `(disambiguation)` in English and its equivalent in the other languages is removed.

```
<transDict>
<namedEntity id='2134'>
  <an>Organizazión d'as Nazions Unitas</an>
  <ar>الأمم المتحد</ar>
  <bs>Ujedinjeni narodi</bs>
  <el>Οϱγανισμός Ηνωμένων Εθνών</el>
  <ga>Nisiin Aontaithe</ga>
  <gl>ONU</gl>
  <he>תודחואמה תומואה</he>
  <hu>Egyesült Nemzetek Szervezete</hu>
  <lb>Vereent Natiounen</lb>
  <nds>Vereente Natschonen</nds>
  <ru>Организация Объединённых Наций</ru>
  <tr>Birle§mi§ Milletler</tr>
  <en>United Nations</en>
  ...
  <kr>국제 연합</kr>
  <jp>国際連合</jp>
  <zh>联合国</zh>
</namedEntity>
</transDict>
```

Figure 4: Translation and Transliteration dictionary entry for the named entity *United Nations*.

3. for each Redirect Page r_j which links to a NE ne_l: update *DisDict* with an entry that contains r_j*.title* as key and ne_l as its value.

Thus, the *DisDict* represents relations between proper names and the corresponding NEs they may denote. (Bunescu, 2007; Zesch et al., 2007) These relations can be of one or more different types. Among these are:

- lexical relations, such as synonymy (e.g. the redirect from `Charles Chaplin` to `Charlie Chaplin`), meronymy (e.g. the redirect from `Phil Stanton` to `Blue Man Group`) or polysemy and homonymy respectively (e.g. `Chaplin`, which refers to people (e.g. `Charlie Chaplin`), to films (`Chaplin (1992 film)`) or to places (`Chaplin,Connecticut`))

- abbreviations, such as acronyms (e.g. the disambiguation page `ACC`, which may, among others, refer to `ACCBank`, `Asian Cricket Council` or `American College of Cardiology`)

- spelling variations, including misspellings (e.g. the redirect from `Cristopher Lee` to `Christopher Lee`) or alternative spellings.

Synonymy, meronymy and other spellings are covered by Redirect Pages, whereas polysemy and homonymy are covered by Disambiguation Pages. Abbreviations are covered by both, Redirect and Disambiguation Pages.

Within the *DisDict*, the *RedirectDict* (cf. step 3 above) is distinguished from the rest of the dictionary. This is necessary in order to be able to correctly identify and process links of the type `[[redirect]]`, where `redirect` is a redirect to a NE, which can't therefore be directly inferred from the link itself.

3.2.2. Creating the Context Dataset

Let l_{ne_j} be a link with the surface form *srf* and the target *tar*, where *tar* is a unique NE ne_j. *tar* can be given within the link, i.e. l_{ne_j} has the form `[[tar|srf]]` or just `[[tar]]`, or it is yielded by the *RedirectDict*, i.e. l_{ne_j} is of the form `[[srf]]` and ne_j is the corresponding value of the key *srf* in the *RedirectDict*.

The *Context Dataset* comprises all contexts of link occurrences of the type l_{ne}, i.e. contexts of all occurrences of NEs. These contexts are extracted from Wikipedia articles. The NEs are disambiguated as determined by the linking structure. In order to allow for a wide range of usages of the dataset, the context window of a NE ne_j is set to the entire paragraph of the Wikipedia article in which ne_j occurs.

In addition, the complete *DisDict* serves the creation of a *Context Dataset* which comprises solely contexts of ambiguous occurrences of NEs: An ambiguous occurrence is present if the *DisDict* contains a link l_{ne_j}'s surface form *srf* and there are at least two NEs l_{ne_j} can denote, namely the corresponding values of *srf* in the *DisDict*, one of which being ne_j (cf. (Bunescu and Pasca, 2006)).

The *Context Dataset* is created for each language. An example of the constructed dataset for Chinese and English can bee seen in Figure 5.

```
<dataset  neID='2134' lang='zh'
          neStr='联合国'>
<context id='1'>
  <surfaceForm>联合国</surfaceForm>
  <leftContext>
  联合国总部大楼，也称为“联合国大厦”，是
  </leftContext>
  <rightContext>
  总部所在地，在美国纽约市曼哈顿区的東側，
  可以俯瞰東河
  </rightContext>
</context>
...
</dataset>
<dataset  neID='2134' lang='en'
          neStr='United Nations'>
<context id='0'>
  <surfaceForm>United Nations</surfaceForm>
  <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>
...
```

Figure 5: Excerpts from the Chinese and English context dataset for the NE *United Nations* (联合国)

4. Evaluation

The performance of our translating approach depends on two main factors: the quality of the initial NE list ac-

Annotation Set Overview

Set	Markables	Annotators	Kappa
1	2000	2	0.774
2	2000	3	0.771

First Annotation Set Results

Annotator	True Positives	False Positives	Precision
1	1900	100	0.950
2	1924	76	0.962

Second Annotation Set Results

Annotator	True Positives	False Positives	Precision
1	1872	128	0.936
2	1911	89	0.956
3	1914	86	0.957

Pairwise Annotator Agreement

Annotator	1	2	3
1	1.000	0.734	0.922
2	0.734	1.000	0.684

Table 1: Evaluation Results

quired in the source language, and the density of the cross-language links along which we translate the NEs to the target languages. We therefore evaluated our system with regard to (i) precision of our NER heuristic for the English Wikipedia and (ii) the increase in coverage yielded by our triangulation method. For precision of the acquired target language NEs we rely on the quality of the cross-language links manually coded in Wikipedia.

4.1. Precision of the NER heuristics

4.1.1. Annotation

In order to estimate the actual precision of the heuristics utilised by our system to recognise NEs we manually annotated two sets of strings recognised by our system as NEs. Both sets consisted of 2000 examples and were annotated by two and three annotators respectively.

We asked our annotators to adhere to the annotation guidelines[22] for the CoNLL-2003 language independent NER shared task (Sang and Meulder, 2003).

Annotator Agreement

To measure the difficulty of the task of filtering NEs from Wikipedia titles, we calculated the inter-annotator agreement in terms of Cohen's Kappa for the first annotation set and pairwise Cohen's Kappa and Fleiss' Kappa for the second. See table 1 for an overview. Kappa values are computed according to (Cohen, 1960; Fleiss, 1971).

The inter-annotator agreement of 0.774 on the first anotation set and 0.771 on the second, confirms our initial assumption that recognising NEs is a challenging task, even if annotators are given annotation guidelines. We would like to point out that the Fleiss Kappa value on the second set stems mainly from the fact that one of the three annotators is in disagreement with the other two. This is reflected in values of 0.734 and 0.684 for the pairwise Cohen's Kappa for the former and an exceptionally high value of 0.922 for the latter.

There were a total of 38 disagreements between the two annotators of the first set, of which 36.84 % are due to the fact that strings denoting biological classifications were handled differently and 21.05 % were obvious annotation errors due to inconsistencies with the annotation guidelines. Disagreements in the other cases were mainly due to strings that do not unambiguously denote a single entity, but rather a class or group of multiple entities with specific features.

4.1.2. Results

Using (Bunescu and Pasca, 2006)'s heuristic we extracted a total of 1,547,586 NEs from the English Wikipedia[23]. The heuristic achieved an average precision of 0.95, from which we conclude that the set of NEs we extracted from the English Wikipedia indeed builds a highly reliable seed for our translation step.

Nevertheless, an improvement of the NER for English is desirable. Analysing the false positives agreed upon by at least two annotators shows a promising direction to do so: 75.32 % of these errors arise because article titles like *History of the Caribbean*, *Politics of the People's Republic of China* and *Geography of Egypt* are erroneously recognised by the heuristic as being a NE. This type of error can be easily dealt with by identifying typical phrases such as *History of* which, if encountered in a given string, will cause it to be classified as not being a NE.

4.2. Multilingual Approach: Coverage by triangulation

We applied the triangulation to the 13 largest language versions of Wikipedia in number of articles (cf. figure 2), as well as to the Turkish and the Swahili Wikipedia. The number of NEs acquired for each language by triangulation is given in table 2. In total, 77,694 additional NE translations could be acquired, corresponding to a 4.5% increase in coverage. The Turkish, Spanish, Italian and Russian languages profited most from the computation of transitive closure, with up to 7.7% increase.

4.3. Multilingual Approach: Context Datasets

For English and for each of the languages mentioned in subsection 4.2., a *DisDict* and a *Context Dataset* were created. Table 3 shows the number of NE contexts extracted from the Wikipedia dumps for each language. The third column presents additional figures for the number of contexts that pertain to ambiguous NEs. Keep in mind that these occurrences have been disambiguated as determined by the linking structure. These contexts can serve as training data for NED classifiers. Moreover, we report the average number of contexts for all ambiguous NEs of one language in terms of mean and median.

Table 4 shows the distribution of the sizes of extracted context sets per NE, for the English *Context Dataset*. For four NEs, more than 100,000 contexts could be extracted. These are: `United States` (371,706 contexts),

[22] `http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt`

[23] XML Database dump enwiki-20080103-pages-articles.xml

Language	**Initial**	**Final**	**Difference**	
			absolute	**%**
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. The *Initial* column shows the number of NEs acquired in the target languages by extracting the cross-language links from the English Wikipedia. The *Final* column shows the number of NEs after having applied triangulation.

Language	**Number of contexts**			
	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 3: Number of contexts extracted from the different Wikipedia language versions.

England (122,924), United Kingdom (114,140) and Germany (100,717). We extracted 23,874 NEs with more than 200 contexts, i.e. 1.5% of all NEs correspond to 44% of all English contexts.

Figure 6 gives an overview of the extracted number of contexts per NE for the other languages.

5. Conclusion and Outlook

We have presented a method to build a huge multilingual resource for NE recognition and NE disambiguation by acquiring translations of NEs from the Wikipedia encyclopedia. We extracted NEs for one source language and acquired translational equivalents for them in many target languages, making use of Wikipedia's cross-language links. We built a comprehensive database of NEs that comprises unambiguously translated and transliterated NEs, as well as ambiguous proper names with corresponding disambiguation dictionaries for the NEs the ambiguous proper names can denote. For all NEs, our resource provides language-specific context datasets of disambiguated NEs. These can serve as annotated training data for NER and NED systems.

	Number of contexts	**Number of NEs**
>=	100,000	4
>=	7,500	235
>=	5,000	181
>=	2,500	585
>=	1,000	2,108
>=	500	4,439
>=	200	16,326
total		**23,878**

Table 4: Number of contexts per NE in the English *Context Dataset*

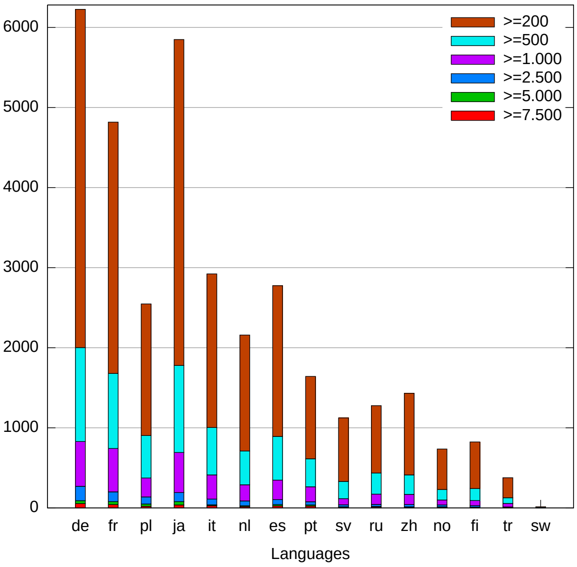

Figure 6: Number of contexts per NE

Up to now, HeiNER does not contain semantic class labels for NEs/NE types. This further extension could be achieved by associating the extracted NEs with the Wikipedia categories their respective pages are labelled with. By linking the NEs to the Wikipedia taxonomy of (Ponzetto and Strube, 2007), suitable class labels could be automatically extracted from the taxonomy. This represents a non-trivial task that we leave for future work.

Nevertheless, due to the Translation Dictionary contained in HeiNER, it is possible to apply one of the well-known NE recognizers available for English, e.g. the Stanford CRFClassifier[24], to the English NEs contained in

[24] The package is available under http://nlp.stanford.edu/software/CRF-NER.shtml, the underlying work being reported in (Finkel et al., 2005).

HeiNER and transfer the class/type information yielded by the NE classifier to the target language of interest. In a second step, linguistic contexts for the resulting set of NEs in the target language can be looked up in HeiNER's context dataset and used to bootstrap other NEs of the same type in the same language. Analogous scenarios for crosslingual NED are also conceivable.

Our resource should be of high value in particular for resource-poor languages. For such languages in particular, NE recognisers can be built or improved quickly, using the context information as training material. HeiNER will be successively extended by applying the methods presented in this paper to further language language versions of Wikipedia.

Our resource can be exploited for many interesting NLP applications and research issues. In future work, we plan to use the language specific datasets of NE contexts to create topic signatures, i.e. families of topically related terms with associated weights (Lin and Hovy, 2000).

HeiNER will be made publicly available at `http://heiner.cl.uni-heidelberg.de`.

6. Acknowledgements

We thank Anette Frank, Philipp Cimiano and Nils Reiter for numerous valuable and very helpful discussions throughout the work reported in this paper.

7. References

Sisay F. Adafre and Maarten de Rijke. 2005. Discovering missing links in Wikipedia. In *LinkKDD '05: Proceedings of the 3rd International Workshop on Link Discovery*, pages 90–97. ACM Press.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the EACL 2006 Workshop on New Text-Wikis and Blogs and other Dynamic Text Sources*.

I. Alegria, N. Ezeiza, and I. Fernandez. 2006. Named entities translation based on comparable corpora. In *Proceedings of the EACL Workshop on Multi-Word-Expressions in a Multilingual Context*, Genoa, Italy.

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy*, pages 9–16, April.

Razvan Bunescu. 2007. *Learning for Information Extraction. From Named Entity Recognition and Disambiguation to Relation Extraction*. Ph.D. thesis, University of Texas at Austin.

Jocob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Andreas Eisele. 2005. First steps towards multi-engine machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 155–158, 6.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into Information Extraction systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 363–370, 6.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Wei Gao, Kam-Fai Wong, and Wai Lam. 2004. Phoneme-based transliteration of foreign names for the oov problem. In *Proceedings of IJCNLP*, pages 110–119.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12:3–23.

Anne-Laure Ligozat, Brigitte Grau, Isabelle Roba, and Anne Vilnat. 2006. Evaluation and improvement of cross-lingual Question Answering strategies. In *Proceedings of the EACL Workshop on Multilingual Question Answering*, pages 110–119.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational linguistics*, pages 495–501, Morristown, NJ, USA.

Rada Mihalcea. 2007. Using Wikipedia for automatic Word Sense Disambiguation. In *Proceedings of the NAACL 2007*, Rochester, April.

David Nadeau, Peter D. Turney, and Stan Matwin. 2006. Unsupervised named entity recognition. Generating gazetteers and resolving ambiguity. In *Proceedings of the 19th Canadian Conference on Artificial Intelligence*, pages 266–277, Quebec.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*, pages 1440–1447, July.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 7th Conference on Natural language Learning at HLT-NAACL 2003*, pages 142–147, Morristown, NJ, USA.

Richard Sproat, Tao Tao, and ChengXiang Zhai. 2006. Named entity transliteration with comparable corpora. In *Proceedings of the 44th Annual Meeting of the ACL*, Sydney.

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1219–1224, Boston.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-lingual Information Retrieval. In *Proceedings of the ACL Workshop on Multilingual and Mixed-Language Named Entity Recognition*, pages 57–64.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. Analyzing and accessing Wikipedia as a lexical semantic resource. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors, *Data Structures for Linguistic Resources and Applications*, pages 197–205. Narr, Tübingen.