Statistical Analyses of Named Entity Disambiguation Benchmarks

Nadine Steinmetz, Magnus Knuth, and Harald Sack

Hasso Plattner Institute for Software Systems Engineering, Potsdam, Germany,
firstname.lastname@hpi.uni-potsdam.de

Abstract. In the last years, various tools for automatic semantic annotation of textual information have emerged. The main challenge of all approaches is to solve ambiguity of natural language and assign unique semantic entities according to the present context. To compare the different approaches a ground truth namely an annotated benchmark is essential. But, besides the actual disambiguation approach the achieved evaluation results are also dependent on the characteristics of the benchmark dataset and the expressiveness of the dictionary applied to determine entity candidates. This paper presents statistical analyses and mapping experiments on different benchmarks and dictionaries to identify characteristics and structure of the respective datasets.

Keywords: named entity disambiguation, benchmark evaluation

1 Introduction

One essential step in understanding textual information is the identification of semantic concepts within natural language texts. Therefore multiple Named Entity Recognition systems have been developed and become integrated in content management and information retrieval systems to handle the flood of information.

We have to distinguish between *Named Entity Recognition* (NER) systems that refer to finding meaningful entities within a given natural language text that are of a specific predetermined type (as e. g., persons, locations, or organizations) and *Named Entity Disambiguation* (NED) systems (sometimes also referred to as *Named Entity Mapping* or *Named Entity Linking*) that take the NER process one step further by interpreting named entities to assign a unique meaning (*entity*) to a sequence of terms. In order to achieve this, first all potential entity candidates for a phrase have to be determined with the help of a dictionary. The number of potential entity candidates corresponds to the level of ambiguity of the underlying text phrase. Taking into account the context of the phrase, as e. g. the sentence where the phrase occurs, a unique entity is selected according to the meaning of the phrase in a subsequent disambiguation step.

Multiple efforts compete in this discipline. But, the comparison of different NED systems is difficult, especially if they don't use a common dictionary for entity candidate determination. Therefore, it is highly desirable to provide common benchmarks for evaluation. On the other hand, benchmarks are applied to tune a NED system for its intended purpose and/or a specific domain, i. e. context and pragmatics of the NED system are fixed to a specific task. To achieve this multiple benchmark datasets have been created to evaluate such systems. To evaluate a NED system and to compare its performance against already existing solutions the system's developer should be aware of the characteristics of the available benchmarks.

In this paper, prominent datasets – dictionary datasets as well as benchmark datasets – are analyzed to gain better insights about both their characteristics and on their capabilities while considering also potential drawbacks. The datasets are statistically analyzed for *mapping coverage*, *level of ambiguity*, *maximum achievable recall*, as well as *difficulty*. All benchmarks and evaluation results are available online to achieve more target-oriented evaluations of NER and NED systems.

The paper is organized as follows: Section 2 gives an overview on NED tools and comparison approaches and introduces the benchmarks and dictionaries utilized in this paper. Statistical information about the benchmarks are presented in Section 3. Experiments using four different dictionaries on three different benchmarks are described and discussed in Section 4. Section 5 concludes the paper and summarizes the scientific contribution.

2 Related Work

Semantic annotation of textual information in web documents has become a key technology for data mining and information retrieval and a key itself towards the Semantic Web. Several tools for automatic semantic annotation have emerged for this task and created a strong demand for evaluation benchmarks to enable comparison. Therefore, a number of benchmarks containing natural language texts annotated with semantic entities have been created. Cornolti et al. present a benchmarking framework for entity-annotation tools and also compare the performances of various systems [3]. This evaluation indicates a difference between several applied datasets, but does not analyze their causes in further detail. Gangemi describes an approach of comparing different annotation tools without the application of a benchmark [5]. The baseline for the evaluation is defined by the maximum agreement of all evaluated automatic semantic annotation tools. Unfortunately, such a baseline does not take into account different semantic annotation levels in terms of the special purposes the evaluated tools have been developed for.

DBpedia Spotlight is an established NED application that applies an analytical approach for the disambiguation process. Every entity candidate of a surface form found in the text is represented by a vector composed of all terms that co-occurred within the same paragraphs of the Wikipedia articles where this entity is linked [9]. The approach has been evaluated on a benchmark containing ten semantically annotated New York Times articles. This benchmark is described in Section 3.1 and part of the presented experiments. DBpedia Spotlight applies a Wikipedia based dictionary – a Lexicalization dataset – to determine potential entity candidates. This dataset is also part of the presented experiments and described in the next section.

AIDA is an online tool for disambiguation of named entities in natural language text and tables [12]. It utilizes relationships between named entities for the disambiguation. AIDA applies a dictionary called *AIDA Means* to determine potential entity candidates. This dictionary is further described in the next section and also under observation for the presented experiments described in Section 4. AIDA has been evaluated on a benchmark created from the CoNLL 2003 dataset[1]. Since this dataset is not available

[1] http://www.cnts.ua.ac.be/conll2003/ner/

for free, *KORE 50* – a subset of the AIDA benchmark dataset – has been used for the experiments in this paper which is described in Section 3.1.

3 Benchmark Dataset Evaluation

3.1 Benchmark Datasets

The benchmark datasets under consideration contain annotated texts linking enclosed lexemes to entities. Based on these benchmarks the performance of NED systems can be evaluated. Within this work, we restrict our selection of benchmark datasets to those containing (a) english language texts (b) originating from authentic documents (e. g. newswire), (c) containing annotations to DBpedia entities or Wikipedia articles, and (d) involving context at least on sentence level.

The DBpedia Spotlight dataset [9] has been created for the eponymous NED tool. It contains 60 natural language sentences from ten different New York Times articles with overall 249 annotated DBpedia entities, i. e. the entities are not explicitely bound to mentions within the texts, which causes a certain lack of clarity. Therefore, we (in all conscience) retroactively have allocated the entities to their positions within the texts. The entities `dbp:Markup_Language` and `dbp:PBC_CSKA_Moscow` could not be linked in the texts, since there was also a more specific entity enlisted occupying their solely possible location, e. g. *hypertext markup language* has been annotated with `dbp:HTML` rather than `dbp:Markup_language`.

KORE 50 (AIDA) [7] is a subset of the larger AIDA corpus [8], which is based on the dataset of the CoNLL 2003 NER task. The dataset aims to capture hard to disambiguate mentions of entities and it contains a large number of first names referring to persons, whose identity needs to be deduced from the given context. It comprises 50 sentences from different domains, such as music, celebrities, and business and is provided in a clear TSV format.

The Wikilinks Corpus [10] has been introduced recently by Google. The corpus collects hyperlinks to Wikipedia gathered from over 3 million web sites. It has been transformed to RDF using the NLP Interchange Format (NIF) by Hellmann et al. [6]. The corpus is divided in 68 RDF dump files, from which the first one[2] has been used for Lexicalization Statistics (cf. Section 4). The intention behind links to Wikipedia articles needs to be considered in a different way compared to the intention of the other two datasets, since links have been created rather for informational reasons. For each annotation the original website is named, which allows to recover the full document contexts for the annotations, though they are not contained in the NIF resource so far. This benchmark cannot be considered as a gold standard. In some cases mentions are linked to broken URLs, redirects or semantically wrong entities. This issue is also discussed in Section 4.

For further processing NIF representations of KORE 50 and DBpedia Spotlight have been created, which are accessible at our website[3]. Further datasets not considered in this paper are e. g. the complete AIDA/CoNLL corpus [8], the WePS (Web people search) evaluation dataset [1], the cross-document Italian people coreference (CRIPCO) corpus [2], and the corpus for cross-document coreference by Day et al. [4].

[2] It can be assumed that the slices are homogeneously mixed.

[3] `http://www.yovisto.com/labs/ner-benchmarks/`

3.2 Benchmark Statistics

The three benchmark datasets under consideration cover different domains, e. g. though all datasets originate from authentic corporas varying portions have been selected and different types of entities have been annotated. Table 1 shows the distribution of DBpedia types within the benchmark dataset.

About 10% of the annotated entities in the DBpedia Spotlight dataset are locations and majority of about 80% of the annotated entities are not associated with any type information in DBpedia. Since the DBpedia Spotlight dataset originates from New York Times articles, the annotations are embedded in document contexts.

Table 1. Distribution of DBpedia types in Benchmark Datasets

Class	Spotlight		KORE 50		Wikilinks	
	entities	mentions	entities	mentions	entities	mentions
total	249	331	130	144	2,228,049	30,791,380
untyped	79.9%	80.1%	18.5%	17.4%	66.5%	60.7%
Activity	<1%	<1%	–	–	<1%	<1%
- Sport	<1%	<1%	–	–	<1%	<1%
Agent	2.4%	2.7%	66.9%	70.8%	18.9%	18.7%
- Organisation	<1%	<1%	18.5%	19.4%	5.3%	5.8%
- - Company	<1%	<1%	9.2%	9.7%	1.8%	1.8%
- - SportsTeam	–	–	7.7%	6.9%	<1%	<1%
- - - SoccerClub	–	–	7.7%	6.9%	<1%	<1%
- Person	2.0%	2.4%	48.5%	51.4%	13.6%	12.9%
- - Artist	–	–	17.7%	18.8%	3.4%	3.5%
- - - MusicalArtist	–	–	17.7%	18.8%	1.8%	1.7%
- - Athlete	–	–	6.9%	8.3%	1.2%	<1%
- - - SoccerPlayer	–	–	5.4%	6.3%	<1%	<1%
- - Officeholder	<1%	<1%	4.6%	4.2%	1.1%	1.2%
Colour	1.6%	1.5%	–	–	<1%	<1%
Disease	1.6%	1.2%	–	–	<1%	<1%
EthnicGroup	1.2%	1.8%	–	–	<1%	<1%
Event	1.2%	<1%	–	–	1.0%	1.5%
Place	10.4%	10.0%	10.8%	10.4%	9.6%	12.2%
- ArchitecturalStructure	2.0%	1.5%	3.1%	2.8%	1.8%	1.6%
- - Infrastructure	1.6%	1.2%	<1%	<1%	<1%	<1%
- PopulatedPlace	7.2%	7.6%	5.4%	5.5%	5.1%	8.0%
- - Country	3.6%	3.3%	–	–	<1%	2.7%
- - Region	<1%	<1%	–	–	<1%	1.0%
- - Settlement	2.4%	3.3%	3.8%	3.5%	3.8%	4.1%
- - - City	1.6%	2.1%	2.3%	2.1%	<1%	1.3%
Work	<1%	<1%	6.2%	6.3%	6.9%	7.3%
- Film	–	–	–	–	1.9%	1.5%
- MusicalWork	<1%	<1%	3.1%	3.5%	1.2%	<1%
- - Album	<1%	<1%	3.1%	3.5%	<1%	<1%
Year	<1%	<1%	–	–	<1%	<1%

The KORE 50 dataset contains 144 annotations which mostly refer to agents (74 times `dbo:Person` and 28 times `dbo:Organisation`). Only a relatively small amount (18.5%) of annotated entities does not provide any type information in DBpedia. The context for the annotated entities in the KORE 50 dataset is limited to (relatively short) sentences.

The by far largest dataset is Wikilinks. Its sheer size allows to extract sub-benchmarks for specific designated domains, e. g. there are about 281,000 mentions of 8,594 different diseases. However, a large amount (66%) of the annotated entities does not provide any type information in DBpedia and the largest amount of the typed entities refer to an agent (18.9%).

4 Lexicalization Statistics and Discussion

The benchmarks described in Section 3.1 are constructed to evaluate NED algorithms. The evaluation results of a NED method are not only dependent on the actual algorithm used to disambiguate ambiguous mentions but also on the structure of the benchmark and the underlying dictionary utilized to determine entity candidates for a mention. A *mention mapping* or *mapped mention* refers to a mention of a benchmark that is assigned to one or more entity candidates of the used dictionary. The following section introduces several dictionaries.

4.1 Dictionary Datasets

Dictionaries contain associations that map strings (surface forms) to entities represented by Wikipedia articles or DBpedia concepts. Typically, dictionaries are applied by NED systems in an early step to find candidates for lexemes in natural language texts. In a further (disambiguation) step the actual correct entity has to be selected from all these candidates.

The DBpedia Lexicalizations dataset [9] has been extracted from Wikipedia interwiki links. It contains anchor texts, the so called surface form, with their respective destination article. Overall, there are 2 million entries in the DBpedia Lexicalizations dataset. For each combination the conditional probabilities $P(uri|surfaceform)$[4], $P(surfaceform|uri)$, and the pointwise mutual information value (PMI) are given. Subsequently, this dictionary is referred to as *DBL* (*DB*pedia *L*exicalizations).

Google has released a similar, but far larger dataset: Crosswiki [11]. The Crosswiki dictionary has been build at webscale and includes 378 million entries. This dictionary is subsequently referred to as *GCW*. Similar to the DBL dataset the probability $P(uri|surfaceform)$ has been calculated and is available in the dictionary. This probability is used for the experiments described in Section 4.2.

The *AIDA Means* dictionary is an extended version of the YAGO2[5] means relation. The YAGO means relation is harvested from disambiguations pages, redirects, and links in Wikipedia [12]. Unfortunately, there is no information given what the extension includes exactly. The AIDA Means dictionary contains ∼18 million entries. Subsequently, this dictionary is referred to as *AIDA*.

[4] The measure is used later on for the experiments as Anchor-Link-Probability (cf. Section 4)

[5] `http://www.yago-knowledge.org/`

In addition to the three already existing dictionaries described above, we have constructed an own dictionary. Similar to the YAGO means relation this dictionary has been constructed by solving disambiguation pages and redirects and using these alternative labels additionally to the original labels of the DBpedia entities. Except the elimination of bracket terms (e. g. the label *Berlin (2009 film)* is converted to *Berlin* by removing the brackets and the term within them) no further preprocessing has been performed on this dictionary. Thus, all labels are presented in original case sensitivity. Further evaluation on this issue is described in Section 4.3. This dictionary is subsequently referred to as *RDM* (*R*edirect *D*isambiguation *M*apping).

4.2 Experiments

To identify several characteristics of the introduced dictionaries as well as consolidate assumptions about the structure of the benchmarks the experiments described in the following sections have been conducted. For performance issues only a subset of the Wikilinks benchmark has been used for the following experiments. For the subset the first dump file containing 494,512 annotations and 192,008 distinct mentions and assigned entities has been used.

Mapping Coverage First, the coverage of mention mappings is calculated. All annotated entity mentions from the benchmarks are looked up in the four different dictionaries. If at least one entity candidate for the mention is found in the dictionary a counter is increased. This measure is an indicator for the expressiveness and versatility of the dictionary.

Entity Candidate Count For all mapped mentions the number of entity candidates found in the respective dictionary is added up. The number of entity candidates corresponds to the level of ambiguity of the mention and can be considered as an indicator for the level of difficulty of the subsequent disambiguation process.

Maximum Recall The list of entity candidates for all mapped mentions is looked up whether the annotated entity (from the benchmark) is included. Only if it is contained in the list, a correct disambiguation is achievable at all. Thus, this measure predicts the maximum achievable recall using the respective dictionary on the benchmark.

Recall and Precision achieved by Popularity For Word Sense Disambiguation (WSD) after determining entity candidates for the mentions a subsequent disambiguation process tries to detect the most relevant entity of all candidates according to the given context. For this experiment the disambiguation process is simplified: the most popular entity among the available candidates is chosen as correct disambiguation. To determine the popularity of the entity candidates three different measures are applied:

- Incoming Page Links of entity candidates
- Anchor-Link-Probability within web document corpus
- Anchor-Link-Probability within Wikipedia corpus

The first measure is a simple entity-based popularity measure. The popularity is defined according to the number of incoming Wikipedia page links. The more links point to an entity the more popular the entity is considered. The Anchor-Link-Probability defines the probability of a linked entity for a given anchor text. Thus, the more often a mention is used to link to the same entity the higher is the Anchor-Link-Probability. This probability has been calculated on two different corpora. For the DBL dictionary this probability has been calculated based on the Wikipedia article corpus and for GCW dataset it has been calculated based on all web documents (cf. Section 4.1). The results of this experiment can be considered as an indicator for the degree of difficulty of the applied benchmark in terms of WSD. A high recall and precision by simply using a popularity measure indicates a less difficult benchmark dataset. If a benchmark contains less popular entities the disambiguation process can be considered more difficult.

4.3 Results & Discussion

The experiments described above are discussed in the following paragraphs. For every experiment a table with the achieved results is given. The tables show the results for the four different dictionaries – represented by the columns – on the three different benchmarks – represented by the rows. For comparison issues, for all dictionaries the number of entries and for all benchmarks the number of distinct mentions and their annotated entities is given. For all results the total numbers as well as proportional respectively an averaged value is given. This facilitates the comparison of benchmarks and dictionaries that are significantly differing in number of annotations and size.

The experiments *mapping coverage*, *entity candidate count*, *maximum recall*, and *recall and precision based on page link popularity* have been also performed using case-insensitive mentions and labels in the four different dictionaries. For comparison, these results are presented in the same tables of the respective experiments as the results of the case-sensitive experiments. Recall and precision based on Anchor-Link-Probability have not been calculated as the probabilities for case-insensitive anchors are not available for the DBL and GCW datasets.

Mapping Coverage

- GCW achieves highest coverage (between 94.67% and 100%) due to largest dictionary containing 378 m. entries and its construction method: anchor texts and linked Wikipedia articles in web documents.
- RDM performs worst with only 25.19% on the Spotlight benchmark due to the lack of preprocessing – all labels are given with capital first letters which is not common in English language except for persons, places, organizations.
- Coverage for RDM increased by 69% (to 94%) when mentions in Spotlight benchmark are looked up in dictionary case-insensitive. Also, for the Wikilinks benchmark the coverage using the RDM dictionary is increased by 16% to 76%. The RDM dictionary consists of mainly case-sensitive labels (as no pre-processing has been performed). Persons, organizations, and places are written with a first capital letter in English language texts. Mentions of entities of those types are found in a case-sensitive dictionary, such as RDM. In contrast, mentions of entities that are not of type person, organization or place, as e. g. *internet* are not found in the dictionary. If a benchmark contains mainly mentions of entities of type person, organization, or place the RDM dictionary achieves a high mapping coverage – as

for the KORE 50 benchmark. Case-insensitive selection must increase the coverage, especially if the benchmark contains entity mentions that are not of type person, organization or place. This assumption is consolidated by the increased mapping coverage for the Spotlight and Wikilinks benchmark and the type information of the mentioned entities in the benchmarks presented in Table 1.

- Overall, the dictionaries perform very well or even best on the benchmarks that have been constructed for the evaluation of their respective applications: DBL – Spotlight, AIDA – KORE 50, and GCW – Wikilinks.

The overall results are depicted in Table 2.

Table 2. Coverage of mentions that are mapped to one or more entities – total count and percentage

Dic BM	DBL		RDM		AIDA		GCW		Mention
	2M entries		10M entries		18M entries		378M entries		Count
Spotlight	235	89%	65	25%	227	86%	258	97%	265
KORE 50	117	90%	129	99%	128	98%	130	100%	130
Wikilinks	107,669	56%	114,443	60%	115,646	60%	170,765	89%	192,008
Experiment with case-insensitive mentions and dictionary labels									
Spotlight	241	91%	249	94%	235	89%	258	97%	265
KORE 50	121	93%	130	100%	130	100%	130	100%	130
Wikilinks	114,278	60%	145,241	76%	128,139	67%	171,941	90%	192,008

Entity Candidate Count

- KORE 50 benchmark is intended to contain mentions that are hard to disambiguate – overall, all dictionaries achieve highest entity count for this benchmark.
- For the Wikilinks benchmark all dictionaries achieve low entity candidate count which shows that real world annotations seem not too hard to disambiguate.
- AIDA dictionary assigns most entity candidates on KORE 50 benchmark as the dictionary is constructed for evaluation on that benchmark and is supposedly enlarged by labels especially for that purpose.
- KORE 50 contains many persons that are mentioned by their first name only. This results in a large number of entity candidates.
- Wikilinks benchmark is annotated very sparsely and only assumed 'important' entities are linked.

Overall results are shown in Table 3.

Maximum Recall

- DBL and RDM do not contain all first names of persons as needed for benchmark KORE 50. Thus, the maximum recall decreases compared to mapping coverage.
- AIDA performs poorly on Spotlight benchmark due to the structure of dictionary. The dictionary contains a large number of persons' first names. Apparently, the dictionary does not reflect labels for entities in manually annotated texts.

Table 3. Amount of entity candidates for all mapped mentions – overall and averaged per mapped mention

Dic BM	DBL		RDM		AIDA		GCW		Mention
	2M entries		10M entries		18M entries		378M entries		Count
Spotlight	1,849	7.9	1,024	15.8	6,487	28.6	134,493	521.3	265
KORE 50	2,980	25.5	16,936	131.3	74,967	585.7	36,772	282.9	130
Wikilinks	188,748	1.8	244,977	2.1	299,193	2.6	1,346,446	7.9	192,008
Experiment with case-insensitive mentions and dictionary labels									
Spotlight	3,400	14.1	6,508	26.1	13,336	56.7	367,698	1425.2	265
KORE 50	3,079	25.4	16,946	130.4	75,326	579.4	46,244	355.7	130
Wikilinks	207,181	1.8	145,241	2.1	352,107	2.7	1.8 m.	10.6	192,008

- For RDM dictionary the maximum recall increases by 10% respectively 63% for the two benchmarks Wikilinks and Spotlight, if mentions are looked up case-insensitive. This is a reflection of the structure of the benchmarks and the increased coverage of mapped mentions.
- For the Wikilinks benchmark the maximum achievable recall is low compared to the other two benchmarks. This results from the fact that this benchmark cannot be considered as a gold standard (cf. Section 3.1). If a mention is annotated with a wrong entity there is a high probability that this entity is not contained in the lists of entity candidates.

Overall results are shown in Table 4.

Table 4. Maximum achievable recall – coverage of annotated entities (in the benchmark) for mentions contained in the list of candidates

Dic BM	DBL		RDM		AIDA		GCW		Mention
	2M entries		10M entries		18M entries		378M entries		Count
Spotlight	223	84%	60	23%	63	24%	241	91%	265
KORE 50	87	67%	93	72%	112	86%	110	85%	130
Wikilinks	82,338	43%	86,555	45%	82,565	43%	129,449	67%	192,008
Experiment with case-insensitive mentions and dictionary labels									
Spotlight	224	85%	228	86%	75	28%	242	91%	265
KORE 50	89	68%	93	72%	112	86%	110	85%	130
Wikilinks	86,955	45%	106,713	56%	92,824	48%	130,335	68%	192,008

Recall and Precision achieved by Popularity – Incoming Wikipedia Page Links of Entity Candidates

- Notably GCW performs poorly on all benchmarks compared to maximum achievable recall due to a high entity candidate count. Apparently entity candidate lists often contain more popular but incorrect entities.
- In the KORE 50 benchmark, due to many annotated first names, entity candidate lists contain many prospective entities and apparently the correct candidate is often

not the most popular one compared to the other candidates. This explains the poor performance of all dictionaries on the KORE 50 using page link popularity.
- Compared to the maximum achievable recall (of all dictionaries) on the KORE 50 the achieved recall is very low using a popularity measure as simplified disambiguation process. This confirms the intention of the benchmark to contain mentions that are hard to disambiguate.

Overall results are shown in Table 5.

Table 5. Recall and Precision, if most popular entity – based on *incoming Wikipedia page links* – is mapped to mention

Dic BM		DBL		RDM		AIDA		GCW		Mention Count
Spotlight	R	149	56%	50	19%	36	14%	27	10%	265
	P		63%		77%		16%		10%	
KORE 50	R	49	38%	50	38%	56	43%	20	15%	130
	P		42%		39%		44%		15%	
Wikilinks	R	77,583	40%	81,259	42%	75,104	39%	90,458	47%	192,008
	P		72%		71%		65%		53%	
Experiment with case-insensitive mentions and dictionary labels										
Spotlight	R	129	49%	154	58%	43	16%	26	10%	265
	P		54%		62%		18%		10%	
KORE 50	R	50	38%	50	38%	56	43%	18	14%	130
	P		41%		38%		43%		14%	
Wikilinks	R	81,424	42%	100,179	52%	83,949	44%	85,805	45%	192,008
	P		71%		69%		66%		50%	

Recall and Precision achieved by Popularity – Anchor-Link-Probability in web document corpus

- In general, this popularity based on mention and mapped entity performs better than popularity only based on the entities' incoming Wikipedia page links.
- Especially, the recall of GCW dictionary is increased between 13% and 55%. The increase of the recall for the RDM and AIDA dictionaries are not significantly compared to page link popularity.

Recall and Precision achieved by Popularity – Anchor-Link-Probability in Wikipedia corpus

- For the Spotlight and Wikilinks benchmarks this popularity measure achieves higher recall and precision than the popularity measure provided by GCW dictionary. Probably this results from the fact that the Wikipedia corpus is composed by experienced authors and linked texts are well considered.

Overall results are shown in Table 7.

Table 6. Recall and Precision, if most popular entity – based on *Google popularity score* for mention as anchor for entity – is mapped to mention

Dic BM		DBL		RDM		AIDA		GCW		Mention Count
Spotlight	R	199	75%	55	21%	51	19%	187	71%	265
	P		85%		85%		22%		72%	
KORE 50	R	50	38%	56	43%	59	45%	40	31%	130
	P		43%		43%		46%		31%	
Wikilinks	R	79,235	41%	83,079	43%	78,638	41%	120,225	63%	192,008
	P		74%		73%		68%		70%	

Table 7. Recall and precision, if most popular entity – based on *Spotlight popularity score* for mention as anchor for entity – is mapped to mention

Dic BM		DBL		RDM		AIDA		GCW		Mention Count
Spotlight	R	200	75%	53	20%	51	19%	205	77%	265
	P		85%		82%		22%		79%	
KORE 50	R	36	28%	37	28%	43	33%	43	33%	130
	P		31%		29%		34%		33%	
Wikilinks	R	79,226	41%	82,469	43%	78,073	41%	119,925	62%	192,008
	P		74%		72%		68%		70%	

General Findings

- For a simplified disambiguation process the Anchor-Link-Popularity performs better than Page-Link-Popularity. Anchor-Link-Popularity calculated on the Wikipedia corpus performs better than the measure calculated on the web document corpus.
- Dictionaries perform best on the benchmark constructed for the evaluation of the dictionaries' applications.
- Compared to the maximum achievable recall (of all dictionaries) on the KORE 50 benchmark the achieved recall is very low using a popularity measure as simplified disambiguation process. This confirms the intention of the benchmark to contain mentions that are hard to disambiguate.
- DBL performs very good over all benchmarks, especially using its popularity measure. Taking into account its size (2.2 m. entries) compared to GCW dictionary (378 m. entries) this is a surprising discovery.
- The DBL popularity measure has been calculated based on the linked Wikipedia articles within the Wikipedia article corpus. Most of the Wikipedia articles have been composed by experienced authors who know how to write and distribute links within the corpus. This could be an explanation why the Wikipedia based Anchor-Link-Probability performs better than the popularity based on web documents.

5 Conclusion

Evaluation results of NED approaches are dependent on the structure of the used benchmark dataset as well as on the dictionary used for entity candidate determination. The objective of this paper is to point out the differences of several benchmarks and dictionaries for NED. For this purpose three different benchmarks have been analyzed. Two of them first have been converted into NIF representations and made available online. The analyses included simple statistical information as well as type information of contained entities about the benchmarks. Additionally, four different dictionaries have been applied to determine entity candidates in the benchmarks. Based on our evaluation, important assumptions about the benchmarks have been consolidated and new insights into the characteristics of evaluated benchmarks as well as on the expressiveness of the dictionaries have been delivered. By making all benchmarks and evaluation results available online, evaluation of new NER or NED tools can be achieved more target-oriented with more meaningful results.

References

1. J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. WePS-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
2. L. Bentivogli, C. Girardi, and E. Pianta. Creating a gold standard for person crossdocument coreference resolution in italian news. In *Proc. of the LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, page 19, Marrakech, Morocco, May 2008.
3. M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 249–260, Geneva, Switzerland, 2013.
4. D. Day, J. Hitzeman, M. L. Wick, K. Crouch, and M. Poesio. A corpus for cross-document co-reference. In *Proc. of the LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, May 2008.
5. A. Gangemi. A comparison of knowledge extraction tools for the semantic web. In *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 351–366. Springer Berlin Heidelberg, 2013.
6. S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating NLP using linked data. In *Proc. of 12th Int. Semantic Web Conf.*, Sydney, Australia, October 2013.
7. J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. KORE: Keyphrase overlap relatedness for entity disambiguation. In *Proc. of the 21st ACM international conference on Information and knowledge management*, pages 545–554. ACM, 2012.
8. J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
9. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: shedding light on the web of documents. In *Proc. of the 7th Int. Conf. on Semantic Systems (I-Semantics)*, 2011.
10. S. Singh, A. Subramanya, F. Pereira, and A. McCallum. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015, University of Massachusetts Amherst, 2012.
11. V. I. Spitkovsky and A. X. Chang. A cross-lingual dictionary for english Wikipedia concepts. In *Proc. of the Eight Int. Conf. on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012.
12. M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum. AIDA: an online tool for accurate disambiguation of named entities in text and tables. *PVLDB*, 4(12):1450–1453, 2011.