GerNED: A German Corpus for Named Entity Disambiguation

Danuta Ploch, Leonhard Hennig, Angelina Duka, Ernesto William De Luca, Sahin Albayrak

DAI-Labor, Technische Universität Berlin
Berlin, Germany
{danuta.ploch,angelina.duka,sahin.albayrak}@dai-labor.de
{leonhardhennig,ernesto.deluca}@googlemail.com

Abstract

Determining the real-world referents for name mentions of persons, organizations and other named entities in texts has become an important task in many information retrieval scenarios and is referred to as Named Entity Disambiguation (NED). While comprehensive datasets support the development and evaluation of NED approaches for English, there are no public datasets to assess NED systems for other languages, such as German. This paper describes the construction of an NED dataset based on a large corpus of German news articles. The dataset is closely modeled on the datasets used for the Knowledge Base Population tasks of the Text Analysis Conference, and contains gold standard annotations for the NED tasks of Entity Linking, NIL Detection and NIL Clustering. We also present first experimental results on the new dataset for each of these tasks in order to establish a baseline for future research efforts.

Keywords: German-Language Corpus, Named Entity Disambiguation, Cross-Document Coreference Resolution

1. Introduction

Extracting information from unstructured texts is an important step towards the automatic creation of meaningful content representations, and is crucial in many areas like information retrieval, topic detection and tracking, and knowledge base population. Named Entity Disambiguation (NED) is such an information extraction task, where the goal is to determine the real-world referents of name mentions in text (Bunescu and Pasca, 2006). It is related to word sense disambiguation (Navigli, 2009) and cross-document co-reference resolution (Bagga and Baldwin, 1998), but focuses on the disambiguation of named entities such as persons, organizations, and geopolitical entities.

NED systems typically address two main tasks, Entity Linking (Cucerzan, 2007) and NIL Clustering (Artiles et al., 2010). Entity Linking requires to accurately associate name mentions found in text to predefined entries of a reference knowledge base (KB), and to recognize mentions referring to entities not covered by the KB (NIL detection) (Dredze et al., 2010). As a step towards populating a reference KB with new entries, the goal of the NIL Clustering task is to group together name mentions of NIL queries referring to the same entity.

NED is challenging since name mentions may be ambiguous, and entities can be referenced by different name variants. For example, the name 'Michael Jordan' can refer to the basketball player, but also to a researcher, as well as to many other people not covered by the reference KB. On the other hand, the basketball player Michael Jordan can be designated by his nickname 'Air Jordan' or simply by his family name 'Jordan'.

The development and evaluation of NED approaches require suitable corpora addressing these challenges and covering a wide range of entities of different entity types. Furthermore, similar to most tasks that deal with processing natural language text, it is desirable to develop and evaluate NED methods that work well across different languages, and that account for language-specific differences and cross-lingual similarities. As the effort of constructing such resources is substantial, there are currently only very few larger NED corpora: the resources used in the Knowledge Base Population (KBP) track of the Text Analysis Conference (TAC) (Simpson et al., 2010), and the corpora created for the Web People Search (WePS) challenges (Artiles et al., 2010). However, both challenges focus on English source material, and to the best of our knowledge, there is no comparable corpus for German-language NED.

Our Contributions: In this paper, we introduce GerNED, a German dataset for NED that consists of more than 2,400 confusable name mentions found in a large corpus of German news articles, and uses a reference KB derived from the German version of Wikipedia. We describe the annotation procedure in Section 2, and outline the characteristics of the corpus in terms of entity distribution and confusability in Section 3. In Section 4 we describe an approach that uses standard NED algorithms for the tasks of Entity Linking and NIL Clustering. Finally, we present first experimental results on the new corpus for each of the NED tasks in order to establish a baseline for future research efforts in Section 5. The corpus will be available to the community from the authors upon request.

2. Resource Creation

In this section we describe the structure and the desired qualities of the corpus, as well as the resources used. Then, we discuss our approach to selecting entities and creating queries.

2.1. Structure of the Corpus

Following the structure of the TAC-KBP evaluation datasets, we created a German dataset for NED that consists of evaluation queries, gold-standard answers, a reference knowledge base, and a source document corpus (see Figure 1).

Evaluation queries are specified by

- a *query id*,
- a *surface form* (name mention of an entity) and
- a *source document id*.

The *surface form* corresponds to a text string in the source document referring to a person (PER), organization (ORG), geopolitical entity (GPE) or an entity of an unknown type (UKN). The *source document id* is the file name of the news article containing the surface form. For Entity Linking, a system has to link each query to the correct knowledge base entry, or decide that the query does not have a corresponding entry in the KB. For NIL Clustering, a system must provide a distinct NIL id for each set of co-referent NIL queries. We created gold-standard answers that map each query either to a unique entity id from the reference KB, or, in the case of NIL queries, to a distinct NIL id for each unique entity. Example queries and gold-standard answers are shown in Figure 1.

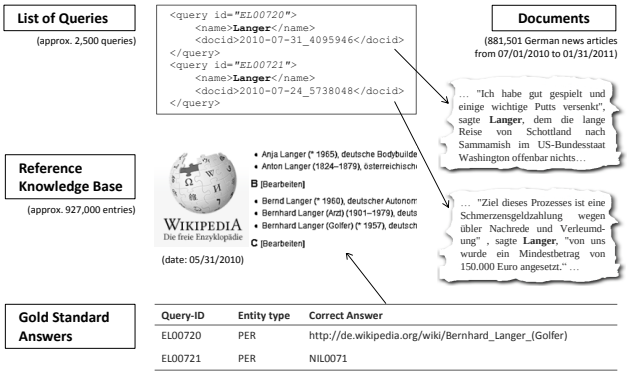

Figure 1: GerNED dataset components. The dataset provides queries of name mentions in German news documents. The gold standard answers link queries to the correct entity id of the reference knowledge base or to a unique NIL id.

2.2. Goals

There are several desirable qualities for NED datasets, as outlined by Simpson et al. (2010), which we considered during the construction of our corpus. Selected entities should be confusable, i.e. they should share a name with another another entity (ambiguity) or be referred to by a set of different name variants, such as spellings, nick names or acronyms (synonymy). Ji et al. (2011) then define the overall confusability of an NED dataset as:

$$ambiguity = \frac{\#surface\ forms\ referring\ to\ more\ than\ one\ entity}{\#surface\ forms}$$

$$variety = \frac{\#entities\ expressed\ by\ more\ than\ one\ surface\ form}{\#entities}$$

Furthermore, the dataset should cover different entity types (PER, ORG, GPE, and UKN), and contain entities with varying mention frequency to cover popular and unpopular entities in the source document corpus. The source document corpus should contain sufficient occurrences of these entities and their name variants. Another requirement is to have a sufficient number of NIL queries exhibiting similar characteristics of name variance, ambiguity and mention frequency as KB queries. In order to allow for comparative evaluations with existing corpora, we modeled the entity distribution and confusability of our dataset on the English TAC-KBP datasets (Simpson et al., 2010; Ji and Grishman, 2011).

2.3. Source Data

We created the source document collection with the friendly support of Neofonie GmbH by crawling web documents from more than 500 German news sources over a time period of seven months from 07/01/2010 to 01/31/2011. News sources include national newspapers and magazines, local newspapers, and news agency feeds. They cover a wide range of news categories and genres, such as politics, financial news, and sports bulletins, and range in length from multi-page essays to brief stock market bulletins. The time period for crawling the newswire documents was chosen to be close to, but later than the 05/2010 epoch of the knowledge base to increase the likelihood of being able to annotate entities not yet covered by the Wikipedia-derived KB.

The raw web documents were transformed into a structured XML format with predefined elements such as title, teaser, and article text. Unwanted elements of the original web document, such as advertisements, navigation menus, footers, etc. were discarded. In total, the corpus contains 881,501 news articles. We assume that a corpus of this size contains enough density and variety to cover a sufficient number of different entities and entity types, contexts of varying difficulty, as well as popular and unpopular entities (Simpson et al., 2010).

2.4. Knowledge Base

The reference knowledge base was constructed from the German Wikipedia. As discussed by Simpson et al. (2010), using Wikipedia has the advantage that Wikipedia entries cover many newsworthy entities. This facilitates choosing candidate evaluation entities as they will very likely be represented in a large corpus composed mainly of newswire articles. We parsed a snapshot of the German Wikipedia from 05/31/2010, and removed all disambiguation, redirect and other meta pages. The resulting reference KB contains approximately 927K entries, as it includes not only named entities, but also general encyclopedia entries.[1] In this respect, our KB differs from the one used in the TAC-KBP evaluations, which consists only of Wikipedia pages having infoboxes. We did not generate separate entity identifiers for each KB entry, but instead re-use the (unambiguous) Wikipedia URI of an entry. In addition, we stored the nor-

[1]Non-entity entries of the KB are ignored for the purpose of entity selection during query creation.

malized page title, as well as raw and cleaned-up versions of the article's text.[2]

2.5. Entity Selection

To model the corpus closely on the TAC datasets, we used queries from the TAC-KBP 2010 training dataset as seeds.[3] First, we translated the English surface forms to German using Wikipedia interlanguage links, since the interlanguage links seem to be a reliable source for translations. Where no interlanguage link was available, we kept the English surface form or we translated name parts that are common nouns by using a dictionary. Out of the different entity types, we had to translate geopolitical entities (GPE) most frequently, e.g. 'Australia' to 'Australien', 'Bavaria' to 'Bayern' and so on, since GPEs often have language-dependent proper names. Usually, it wasn't necessary to translate proper names of persons and organizations since the names remain the same in both language. In some cases though a translation to German provided additional surface forms. For example, both the surface form 'Harvard University' and 'Harvard Universität' are used in German news articles and therefore appropriate translations.

After translating the surface forms we checked their availability in the German KB and the source data. Surface forms found in news articles but missing in the KB served as the basis for NIL queries. We decided to substitute or to complement less popular entities not occurring (often enough) in our news corpus with equivalent German ones. For example, we searched for a German comedian equivalent to an American one (Jerry Springer vs. Harald Schmidt, DeGeneres vs. Engelke), a German town for an English one (Lexington vs. Erfurt), etc.

While creating new queries we ensured that German queries had ambiguous surface forms and selected entities with several name variants like acronyms, abbreviations, and spelling mistakes, following the procedure of the TAC-KBP dataset construction process (Simpson et al., 2010). We included popular as well as less popular entities, and focused on a high confusability. Besides queries for persons (PER), organizations (ORG), and geopolitical entities (GPE), we decided to create also queries for entities with an unknown type (UKN) in order to include interesting queries such as TV series sharing their name with persons or locations.

We found the annotation of NIL queries challenging. Especially, the annotation of GPEs proved difficult, since in contrast to persons and organizations almost every GPE found in a news article was covered by the KB. Out-of-KB-GPEs are therefore under-represented in our corpus. To find NIL queries we followed two strategies: On the one hand we annotated less popular entities sharing the same name with a popular entity to augment the confusability of the queries and to reduce the bias towards popular entities. On the other hand we searched for novel entries added after 05/31/2010 to the German Wikipedia, and included them as a query if they occurred in our news corpus.

[2]We keep the raw article text including markup to allow for later parsing, for example to determine an entity's type from its infobox, or to collect the contextual links of an article.

[3]http://nlp.cs.qc.cuny.edu/kbp/2010/

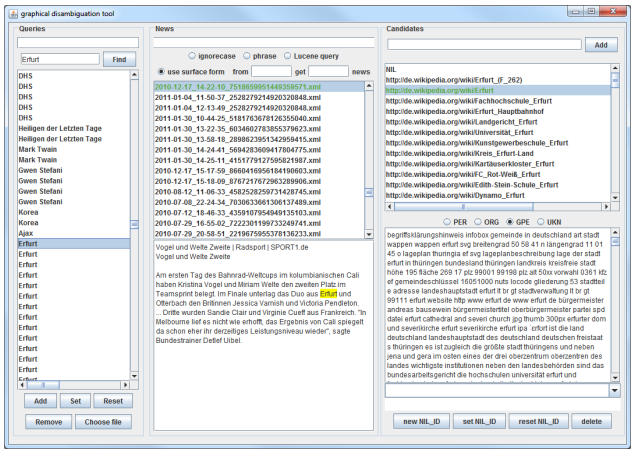

Figure 2: A screenshot of the annotation tool. The selected items form a query.

2.6. Query Creation

We provided an annotation tool (Figure 2) and instructed annotators to create queries by searching for a surface form, selecting a document containing it, and linking it to the correct entity from a list of candidate entities. To create a confusable corpus in terms of ambiguity and variety the annotators were advised to augment the initial surface forms list and if possible to create numerous queries per surface form and entity. They were asked to select different entities per surface form and also to annotate different name variants for one entity using name variants found in Wikipedia or in the news corpus. If a surface form referred to a NIL entity, annotators could create a novel entity or select from the set of previously created NIL entities. Annotators were also advised to specify the entity type and to select only documents containing at least one surface form exactly matching the query. This means, to create e.g. a query containing only a person's family name, the selected document must contain at least one mention of the person by just his or her family name. Altogether, three annotators created the queries. Although each annotator created a different set of queries, all queries were checked later by at least one other annotator. Inconclusive queries were removed from the corpus.

Figure 2 shows a screenshot of the provided annotation tool. The view on the left hand side displays surface forms, the center view lists all documents containing the selected surface form, and view on the right hand side potential candidate entities, as determined by a KB lookup. If an entity is missing in the candidate list, the annotators have the possibility to add the correct KB entry to the list. Working on a NIL query, the tool offers an internal KB with NIL entities so that the annotators either can select an existing NIL entity or add a new entity to the KB of NIL entities. We generate NIL entity identifiers by starting with the id 'NIL0001' and then incrementing the identifiers by one for each new NIL entity. The selected items, as displayed in Figure 2, constitute a single query of the corpus.

	All	KB	NIL
PER	700	450	250
ORG	1127	615	512
GPE	563	542	21
UKN	78	57	21
ANY	2468	1664	804

Table 1: Query distribution of the GerNED corpus

	All	KB	NIL
GerNED corpus	2468	1664	804
TAC-KBP 2010 train	1500	1074	426
TAC-KBP 2010 eval	2250	1020	1230
TAC-KBP 2011 eval	2250	1124	1126

Table 2: Distribution of KB and NIL queries in the GerNED corpus in comparison to the TAC-KBP datasets

3. Corpus Statistics

This section summarizes the key characteristics of the created German NED corpus such as the size, the distribution of KB and NIL queries and of different entity types, and compares them to the TAC-KBP datasets.

The German corpus contains a total of 2468 queries, with 1664 KB and 804 NIL queries. Table 1 shows the distribution of entity types in the corpus for NIL, KB and all queries. The majority of queries (46%) relate to organizations, 28% of the the queries relate to persons and 23% to geopolitical entities. Only 3% are of an unknown type. Altogether, the corpus provides queries for 1190 unique named entities and 1098 distinct surface forms detected in 2417 news articles crawled from 27 German news providers.

The distribution of 70% KB and 30% NIL queries and of different entity types in the GerNED corpus does not reflect necessarily their distribution in news articles, but results in a corpus with an intended focus on confusable queries. Tables 2 and 3 compare the entity distribution of the German dataset with the TAC-KBP datasets of 2010 and 2011.

3.1. Surface Forms

The annotators succeeded in creating numerous queries referencing different entities by the same surface form. On average, each surface form refers to 1.21 entities. Approximately 15% of the surface forms in the corpus are annotated with more than one entity. These queries cover different degrees of difficulty ranging from ambiguous names for entities of the same type (different people named 'Schmidt'), for entities of different types ('Duke Energy' vs. 'Mike Duke') to surface forms used as metonyms – which occur quite often in the created dataset. For example, the surface form 'Erfurt' relates to the town in Thuringia, but may also be used to denote the town's football club. The most ambiguous surface forms in the corpus are 'Duke', 'Erfurt', 'UC', 'MGM', 'San Diego', 'Vancouver', 'Schmidt', 'Weißensee' and 'Justizministerium'.

[5]Ambiguity and variety cannot be computed for the full TAC-KBP 2010 training dataset as its NIL queries are not annotated with distinct NIL identifiers.

	PER	ORG	GPE	UKN
GerNED corpus	700	1127	563	78
TAC-KBP 2010 train	500	500	500	-
TAC-KBP 2010 eval	751	750	749	-
TAC-KBP 2011 eval	750	750	750	-

Table 3: Distribution of entity types in queries of the GerNED corpus in comparison to the TAC-KBP datasets

	All	KB	NIL
GerNED corpus	14.57 %	15.80 %	7.16 %
TAC-KBP 2010 train	N/A	4.12 %	N/A
TAC-KBP 2010 eval	12.93 %	5.70 %	9.31 %
TAC-KBP 2011 eval	13.23 %	12.42 %	7.15 %

Table 4: Ambiguity of the German corpus in comparison to the TAC-KBP datasets[5]

	All	KB	NIL
GerNED corpus	11.09 %	8.39 %	14.90 %
TAC-KBP 2010 train	N/A	3.90 %	N/A
TAC-KBP 2010 eval	1.95 %	2.49 %	1.49 %
TAC-KBP 2011 eval	1.12 %	1.56 %	0.74 %

Table 5: Variety of the German corpus in comparison to the TAC-KBP datasets[5]

Table 4 summarizes the ambiguity of the German corpus as well as of the TAC-KBP datasets for NIL, KB and All queries. Figure 3 shows a detailed overview of the surface form distribution. It illustrates the proportions of surface forms for which one, two, three or more than three named entities are annotated and compares the ambiguity of the surface forms with the TAC-KBP datasets. A surface form is covered on average by 2.3 queries and 64% of the surface forms are annotated in more than one query. Table 4 and Figure 3 show that the ambiguity of the GerNED corpus is comparable to the ambiguity of the TAC-KBP datasets.

3.2. Named Entities

The corpus contains a number of entities denoted by different name variants. On average, each entity is referred to by 1.12 distinct surface forms. Table 5 shows the variety of the German corpus in comparison to the TAC-KBP datasets. It considers the variety of the entire query set as well as of NIL and KB queries and points out that especially the variety of NIL queries is very high. Overall, the annotated queries cover various name variants like acronyms, spelling mistakes and multilingual names. For example, the NATO organization is annotated with the English surface form 'North Atlantic Treaty Organization' and with the German surface form 'Organisation des Nordatlantikvertrags', both occurring in German news articles. Regarding spelling mistakes, the GerNED corpus provides queries such as 'Rotterdamm' for the city in the Netherlands. Some entities are also annotated with their acronyms. For example, the 'Weltgesundheitsorganisation' can be referenced by the acronym 'WHO'. Figure 4 shows that 9.8% of all entities in the German corpus are annotated with two distinct surface forms and 1.3% of the entities are annotated with three surface forms. This usage of synonyms is con-

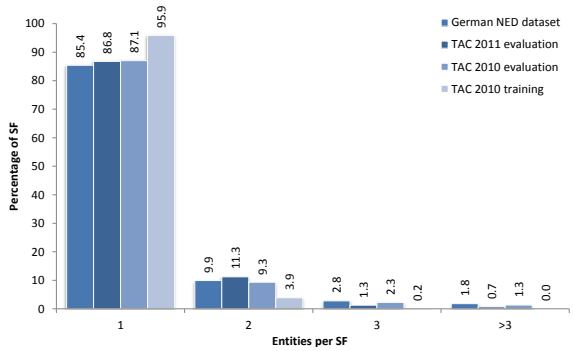

Figure 3: Entities per surface form in comparison to other TAC-KBP datasets

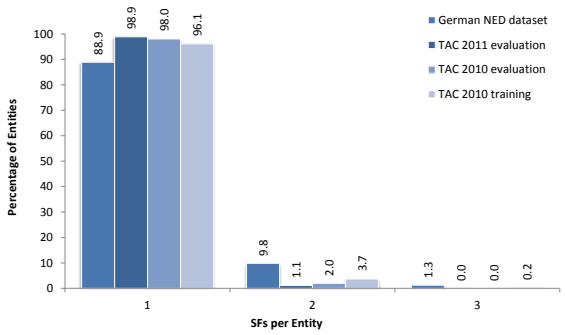

Figure 4: Surface forms per entity in comparison to other TAC-KBP datasets

siderably higher than in the TAC-KBP datasets. The entities with the highest name variety are 'Sido', 'Myanmar', 'Phoenix Hagen', 'Welthandelsorganisation' and 'Harvard University'. Furthermore, there are 2.1 queries per entity on average, and 60% of the entities are annotated in more than one query.

4. Baseline Approach to NED

In this section we describe an approach that uses standard NED algorithms in order to establish a baseline on the presented dataset for future research efforts. We implement the subtasks of Entity Linking and NIL Clustering as a two-step process, first detecting queries referencing a KB entry, and then clustering the remaining NIL queries. This approach follows common practice and is implemented by many systems participating in the TAC-KBP tracks (Ji and Grishman, 2011).

4.1. Entity linking

We formulate Entity Linking as a supervised classification problem. We first generate a set of candidate KB entries for a query, and then rank candidates according to the likelihood that they correspond to the correct entry. Finally, we employ another classification step to detect queries referring to NIL entities. In the following we will briefly describe this approach, for more details we refer the reader to Ploch (2011) and Ploch et al. (2011).

We generate candidates by collecting name variants for each KB entry from article titles, redirect pages, disambiguation pages and the anchor texts of internal Wikipedia links in a preprocessing step. We normalize name variants by lower-casing, and removing punctuation as well as appositives. Candidate generation is then performed by looking up the query name mention in an inverted index mapping name variants to KB entries. This step is geared towards high recall, and prefers a larger candidate set over a smaller one. We limit the candidate set to the N highest scoring results according to the relevance score computed by the index search.[6]

In order to rank candidates, we represent each candidate as a feature vector encoding contextual and KB knowledge as well as comparisons of the two. To provide a realistic baseline, we implement three well-known features which have been shown to be very useful in Entity Linking. The first feature, *surface form popularity (SFP)* is a KB feature that encodes the likelihood with which a particular surface form refers to a given target entity. The entity distribution for a given surface form is determined from the link frequencies of internal Wikipedia anchors, including redirect and disambiguation pages. This feature captures the preference for the "*most frequent sense*" of a name mention (Han and Zhao, 2009). The second feature is based on our use of an inverted index for candidate generation. The *candidate selection score (CS)* measures the relevance score of each KB entity as calculated by the weighted index search, which uses a modified tf-idf weighting scheme over the different parts of a name mention. We found this feature to be very useful in our experiments on KBP 2009 and KBP 2010 datasets, see also Ploch (2011). Our last feature measures the *bag-of-words (BOW)* similarity between the query document and a candidate's KB text using the cosine similarity of tf-idf-weighted word vector representations (Bunescu and Pasca, 2006). We preprocessed document and article texts by performing stemming using Porter's stemmer and removing words occurring in a stop word list.

The NIL detection classifier is based on features derived from the atomic features of all candidates of a given query. We calculate several different features, such as the *max*, *mean*, *min*, *max-mean*, and *max-min*, of the atomic features, using the feature vectors of all candidates of a query.

4.2. NIL Clustering

Our approach for NIL Clustering is based on a hierarchical agglomerative clustering (HAC) algorithm which is a common approach to the task of clustering documents according to the entities they mention (Artiles et al., 2010). The HAC algorithm first assigns each query to its own cluster and then successively merges pairs of clusters until a predefined similarity threshold t is reached or until all queries are assigned to a single cluster. In our baseline scenario we use single-link clustering. We measure the similarity between two queries by calculating the cosine similarity between the tf-idf-weighted word vectors constructed from the document texts of the queries. As for the task of Entity Linking, we first perform stemming and remove stop words before creating the word vectors.

[6] In our experiments, N = 100 was set based on evaluations on the TAC-KBP datasets.

In addition to the HAC baseline approach, we apply three more baseline algorithms to cluster NIL queries. The first two baselines are the straightforward clustering approaches *one-in-one* and *all-in-one* that assign each query to its own cluster or all queries to one single cluster, respectively. Another standard clustering approach for NIL Clustering is to group the queries according to their name mentions. To this end, we lower-case all name mentions and cluster all queries sharing the same name (*all-for-sf*). We provide these baselines to examine whether the created corpus is robust enough to avoid 'cheating'.

5. Evaluation

This section presents experimental results of the baseline approach on the GerNED corpus. We evaluate the steps of Entity Linking and NIL Clustering separately, and additionally conduct an evaluation run considering both steps, where the output of the Entity Linking task is passed as input for the NIL Clustering step.

We measure the quality of this baseline approach using established performance measures adopted in the TAC-KBP Entity Linking task, namely the *micro-averaged accuracy (MAA)* and the *B^3+* metric (Ji and Grishman, 2011). MAA is query-oriented, and measures the fraction of correctly linked queries, whereas B^3+ evaluates the correctness of the clusters of queries referring to the same entity.[7]

5.1. Model Training and Parameter Selection

For Entity Linking, we randomly split the 2468 queries of our dataset into five folds to perform cross-validation. Each split uses 60% of the data for training, 20% for validation, and the remaining 20% of the data for testing. We stratify the folds to ensure a similar distribution of KB and NIL queries, and normalize feature values.

We use a Support Vector Machine classification algorithm (Vapnik, 1995) to train models for candidate ranking and NIL detection, utilizing the LibSVM implementation (Chang and Lin, 2001). For training the candidate ranking classifier we label as a positive example at most one candidate from the set of candidates for a given query, and all others as negative. For training the NIL classifier, we create a single feature vector per query, which we label as positive if the query refers to a NIL entity. Both classifiers use a radial basis function kernel. In each iteration, we perform a grid search to determine optimal values for the SVM's hyperparameters C and γ. The classifier models with optimal performance on the validation data are then used for testing. Results reported in this paper are averaged across the test folds.

We evaluate the baseline NIL Clustering algorithms in two different experimental setups. The first experiment measures the performance when clustering gold-standard NIL queries, in order to avoid a skewed NIL Clustering score resulting from noise introduced by previous NED steps. The second experiment uses the answers predicted by the baseline entity linker to assess the performance of the overall baseline system. We cluster only queries classified as 'NIL' by the entity linker. Queries already linked to the KB are ignored during NIL Clustering but considered for calculating the overall evaluation score.

To tune the threshold parameter t of the HAC algorithm, we randomly split the queries of the dataset into five folds. We ensure that queries for one entity are not distributed across different folds and that the distribution of NIL and KB queries corresponds to their original distribution. To evaluate the HAC approach, we perform cross-validation where we use 20% of the data for finding a good value of the parameter t, and the remaining 80% for testing. The NIL Clustering results are then averaged across the test folds. We evaluate all other NIL Clustering baseline algorithms on 100% of the queries, depending on the evaluation scenario on gold-standard NIL queries or on all queries.

[7]Scorer available at: `http://nlp.cs.qc.cuny.edu/kbp/2011/scoring.html`

	MAA	B^3+ Prec	B^3+ Rec	B^3+ F1
KB	0.627	0.597	0.584	0.590
NIL	0.910	0.765	0.755	0.758
ALL	0.719	0.643	0.639	0.641

Table 6: B^3+ scores and micro-averaged accuracy for the NED baseline system on the GerNED dataset.

	MAA	B^3+ Prec	B^3+ Rec	B^3+ F1
PER	0.744	0.731	0.663	0.695
GPE	0.760	0.747	0.727	0.737
ORG	0.712	0.615	0.610	0.612
UKN	0.294	0.291	0.237	0.261

Table 7: B^3+ scores and micro-averaged accuracy for the NED baseline system by entity type, on the GerNED dataset.

5.2. Results

We present the results of our baseline Entity Linking system in Tables 6 and 7. Table 6 shows the micro-averaged accuracy (MAA) and B^3+ scores for all queries, KB queries only, and NIL queries only. The baseline system achieves an MAA score of 0.719 and a B^3+ F1 score of 0.641 when considering all queries. MAA and B^3+ scores for NIL queries are significantly higher than for KB queries. The better performance on NIL queries suggests that the chosen features are good indicators for discriminating between entities known respectively unknown to the KB, but do not always result in a correct ranking of candidate entities.

In Table 7, we give detailed performance statistics of the baseline NED approach for different entity types. From the table, we see that MAA scores are quite similar for PER, GPE and ORG entities, with GPE and PER entities being slightly easier than ORG entities. Entities of type UKN, however, are much harder to link, and queries for UKN entities only have an MAA score of 0.294. Again, B^3+ scores reasonably mirror MAA scores, with the lower performance for ORG entities (compared to PER, GPE) somewhat more evident. The results shown in this table suggest that the entities selected for the dataset are of a similar difficulty across entity types, with the exception of the much harder queries for UKN entities.

Table 8 summarizes the B^3+ F1 results of the four base-

	all-in-one	one-in-one	all-for-sf	HAC
NIL	0.006	0.719	0.895	0.843
All	0.501	0.615	0.654	0.641

Table 8: Bˆ3+ F1 scores for NIL Clustering baselines on the GerNED dataset.

	MAA	Bˆ3+ Prec	Bˆ3+ Rec	Bˆ3+ F1
GerNED	0.719	0.643	0.639	0.641
TAC-KBP 2010 eval	0.776	0.638	0.583	0.609
TAC-KBP 2011 eval	0.697	0.637	0.596	0.616

Table 9: Comparison of Bˆ3+ scores and micro-averaged accuracy for the baseline system on different NED datasets.

line approaches to NIL Clustering. The first row lists results obtained performing the clustering algorithms on gold-standard NIL queries. It shows that the *all-in-one* approach with its Bˆ3+ F1 score of 0.006 is not suitable for the dataset. The *one-in-one* approach achieves significantly better results. This can be expected due to the structure of the corpus which consists predominantly of small clusters. On average, a cluster contains 1.8 queries and 75% of the clusters consist of only one or two queries. The best results are achieved by the *all-for-sf* approach. It performs even better than the HAC approach which is part of the baseline system. Still, the Bˆ3+ F1 score of 0.895 offers room for improvement. The second row of the table reports results of the baseline system including Entity Linking and NIL Clustering. The results mirror the trend between the clustering approaches evaluated on NIL queries. Since the majority of the queries can be linked to the KB and the NIL queries are well separable by their surface form, the more important subtask is therefore the Entity Linking step.

In order to establish a context for the results of the NED baseline, Table 9 compares the system's performance on different datasets.[8] Bˆ3+ scores are comparable across the three datasets. MAA scores on the TAC-KBP 2010 eval dataset are higher than on the TAC-KBP 2011 eval dataset, and also higher than on the GerNED corpus. A similar observation was made by Ji et al. (2011), who noted that KBP2011 systems perform generally worse on 2011 data than on 2010 data. Overall, the results shown in Table 9 suggest that the GerNED corpus is of similar difficulty as recent TAC-KBP datasets.

6. Related Work

The most prominent resources for the task of Named Entity Disambiguation are the datasets distributed for the Text Analysis Conference's Knowledge Base Population (TAC-KBP) track (Simpson et al., 2010). These datasets provide training data for Entity Linking, NIL Detection and NIL Clustering for English queries and newswire text. The TAC 2011 evaluation in addition included annotated data for cross-lingual Entity Linking using Chinese queries and source documents together with an English knowledge base (Ji et al., 2011). Recently, Mayfield et al. (2011) presented a dataset for cross-lingual Entity Linking that maps English name mentions to non-English documents. The dataset contains approximately 55,000 queries for documents in 21 different non-English languages.

The task of clustering entities without reference to a knowledge base is also addressed by the WEPS challenges (Artiles et al., 2010), which focus on grouping distinct PER entities referenced in web documents retrieved by querying a search engine for person names.

7. Conclusions

We presented a novel, German-language corpus for the task of Named Entity Disambiguation. The corpus consists of a large set of newswire documents, a Wikipedia-derived knowledge base, and a set of queries for ambiguous name mentions. It provides annotations for the subtasks of linking named entity mentions in documents to the knowledge base, and of clustering name mentions of entities not found in the knowledge base according to their real-world referents. We plan to make this corpus available to the larger research community.

Our analysis shows that our corpus is of similar confusability as the TAC-KBP datasets. It contains a higher fraction of synonymous queries than the TAC-KBP datasets, while being comparable in terms of query ambiguity. The German corpus contains fewer NIL queries because of Wikipedia's extensive coverage of named entities. In particular, novel geopolitical entities were hard to find.

Experiments using well-known baseline algorithms for Entity Linking and NIL Clustering give a micro-averaged accuracy of 0.719, and a Bˆ3+ F1 score of 0.641 on the presented dataset. These figures suggest that the GerNED corpus is of similar difficulty as the TAC-KBP datasets, and leave room for more sophisticated approaches.

In future work we intend to expand the NED corpus and augment the source data with non-news documents such as micro-blogs or public sector information in order to study NED in these contexts.

8. References

Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine, and Enrique Amigó. 2010. WePS-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In *Proc. of CLEF 2010*.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proc. of COLING 1998*, pages 79 – 85.

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proc. of EACL 2006*, pages 9–16.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Silviu Cucerzan. 2007. Large-Scale named entity disambiguation based on Wikipedia data. In *Proc. of EMNLP-CoNLL 2007*, pages 708–716.

[8] The evaluation on the TAC-KBP datasets was conducted in the same manner as on the GerNED dataset, i.e. using cross-validation and averaging results across folds.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proc. of COLING 2010*, pages 277–285.

Xianpei Han and Jun Zhao. 2009. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proc. of CIKM 2009*, pages 215–224.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proc. of ACL-HLT 2011*, pages 1148–1158.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC2011 knowledge base population track. In *Proc. of TAC 2011*.

James Mayfield, Dawn Lawrie, Paul McNamee, and Douglas Oard. 2011. Building a Cross-Language entity linking collection in Twenty-One languages. In *Proc. of CLEF 2011*, pages 3–13.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1–69.

Danuta Ploch, Leonhard Hennig, Ernesto William De Luca, and Sahin Albayrak. 2011. DAI approaches to the TAC-KBP 2011 entity linking task. In *Proc. of TAC 2011*.

Danuta Ploch. 2011. Exploring entity relations for named entity disambiguation. In *Proc. of ACL 2011 (Student Session)*, pages 18–23.

Heather Simpson, Stephanie Strassel, Robert Parker, and Paul McNamee. 2010. Wikipedia and the web of confusable entities: Experience from entity linking query creation for TAC 2009 knowledge base population. In *Proc. of LREC'10*.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer, New York.