Search-based Entity Disambiguation with Document-Centric Knowledge Bases

Stefan Zwicklbauer
University of Passau
Passau, 94032 Germany
stefan.zwicklbauer@uni-passau.de

Christin Seifert
University of Passau
Passau, 94032 Germany
christin.seifert@uni-passau.de

Michael Granitzer
University of Passau
Passau, 94032 Germany
michael.granitzer@uni-passau.de

ABSTRACT

Entity disambiguation is the task of mapping ambiguous terms in natural-language text to its entities in a knowledge base. One possibility to describe these entities within a knowledge base is via entity-annotated documents (document-centric knowledge base). It has been shown that entity disambiguation with search-based algorithms that use document-centric knowledge bases perform well on the biomedical domain. In this context, the question remains how the quantity of annotated entities within documents and the document count used for entity classification influence disambiguation results. Another open question is whether disambiguation results hold true on more general knowledge data sets (e.g. Wikipedia). In our work we implement a search-based, document-centric disambiguation system and explicitly evaluate the mentioned issues on the biomedical data set CALBC and general knowledge data set Wikipedia, respectively. We show that the number of documents used for classification and the amount of annotations within these documents must be well-matched to attain the best result. Additionally, we reveal that disambiguation accuracy is poor on Wikipedia. We show that disambiguation results significantly improve when using shorter but more documents (e.g. Wikipedia paragraphs). Our results indicate that search-based, document-centric disambiguation systems must be carefully adapted with reference to the underlying domain and availability of user data.

CCS Concepts

•**Information systems** → **Retrieval tasks and goals;** *Information systems applications;*

Keywords

Entity Disambiguation, Text Annotation, Knowledge Bases, Experiments

i-KNOW '15, October 21 - 23, 2015, Graz, Austria

ISBN 978-1-4503-3721-2/15/10…$15.00

DOI: http://dx.doi.org/10.1145/2809563.2809618

1. INTRODUCTION

Linked (Open) Data provides huge potential to improve information management processes in different domains. For instance, textual information within structured [9, 1] or unstructured data [3, 24] (addressed in this work) can be linked to concepts in the Linked Open Data (LOD) cloud to improve retrieval, storage and analysis of document repositories. The task of entity disambiguation establishes such links between selected text fragments (surface forms) and candidate meanings, referred to as a knowledge base (KB), and faces the problem of semantic ambiguity.

The creation of a disambiguation system demands the choice of a data set that describes all entities as precisely as possible. These data sets describe entities either intensionally, i.e. through a description, or extensionally, i.e. through instances and usage [12, 23]. Intensional definitions can be understood as a thesaurus or logical representation of an entity, as it is provided by LOD repositories. Extensional definitions resemble information on the usage context of an entity, as it is provided by entity-annotated documents. The authors of [22] model these definitions as an *entity-centric* (intensional representation) or *document-centric* KB (extensional representation).

Entity disambiguation with entity-centric KBs from the LOD cloud has been extensively studied on different domains [16, 18, 23, 24]. In contrast, recent work shows that search-based disambiguation with document-centric KBs attains strong results in the biomedical domain [22, 23]). These search-based approaches can be subdivided into two major parts: First, these algorithms retrieve those documents from a document-centric KB, that contain similar textual content as given by the surface form to disambiguate. Second, the most salient entity is selected from the retrieved documents' entity set. However, an expansion and evaluation of these disambiguation approaches on general knowledge (e.g. Wikipedia) is missing.

In our work we define and tackle the following two crucial and important issues concerning search-based entity disambiguation with document-centric KBs:

1. user data influence, i.e. how do the quantity of annotated entities within documents and the document count influence disambiguation results
2. domain dependence, i.e. do results hold true for general domain data sets

Similar to the work of [22] (and [23]), we focus on the biomedical domain to analyze the influence of user data on disambiguation results. In our second experiment we focus on

the Wikipedia data set and evaluate how our disambiguation approach performs on a general knowledge data set.

Overall, our contributions with reference to search-based, document-centric entity disambiguation are the following:

- We show that the number of documents used to classify entities and the number of annotations in these documents must be well-matched to attain the best result.
- We show that disambiguation accuracy is poor on Wikipedia.
- We show that disambiguation accuracy (significantly) increases when using shorter but more documents (e.g. Wikipedia paragraphs).

The remainder of the paper is structured as follows: In Section 2 we define the issues and model document-centric KBs and user data. Section 3 describes the implementation of our disambiguation approach with document-centric KBs. Section 4 analyzes the data sets which are used in our evaluation. Section 5 presents our parameter experiments in the biomedical domain as well as the general domain experiment on the Wikipedia data set. In Section 6 we review related work. Finally, we conclude our paper in Section 7.

2. ISSUES AND MODELING

First, we define two important issues concerning document-centric KBs, namely user data influence and domain dependence. Second, we introduce how we model document-centric KBs and user data in the context of our work.

2.1 Open Issues of Document-Centric KBs

Issue 1: Disambiguation approaches that apply document-centric KBs can be classified in different categories. For instance, the authors of [3, 7, 15] proposed probabilistic, generative approaches, while Zwicklbauer et al. [22, 23] focused on search-based approaches. However, all these algorithms rely on an extensive amount of annotated entities within underlying documents. When using a search-based approach, as realized in our work, the amount of annotated entities used by a document-centric algorithm depends on two major parameters. These are the number of documents to classify entities and the amount of annotations within these documents.

The question remains how these parameters correlate and separately influence the disambiguation results. In our work we refer to both parameters, the amount of entity annotations within documents and the amount of documents for classification, as **user data influence**.

Issue 2: A top-down view on currently available disambiguation algorithms shows that entity-centric ([2, 16, 18]) and generative, probabilistic document-centric ([3, 7, 15]) approaches perform well on general knowledge. Search-based document-centric approaches as proposed in [22, 23] provide good results in specialized knowledge like the biomedical domain. However, search-based document-centric disambiguation approaches are rather unexplored on general knowledge data sets.

This raises the question of how search-based document-centric disambiguation approaches perform on general knowledge (as available in Wikipedia). In the following we denote this issue as **domain dependence**.

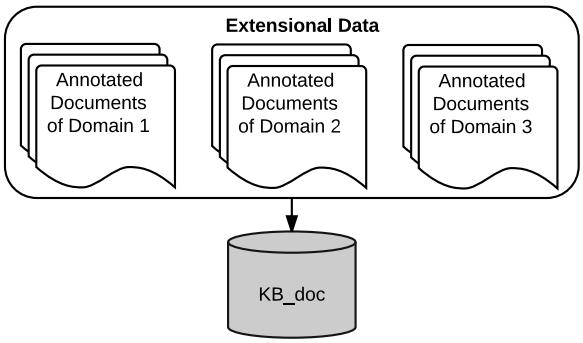

Figure 1: Documents of various data sources are stored within a document-centric KB

2.2 Model

To investigate our two important issues concerning document-centric KBs, we specify and model document-centric KBs as well as user data, which are the core aspects in our work.

2.2.1 Modeling Document-Centric KBs

Entities are described either extensionally or intensionally. In our work we focus on extensional entity descriptions and model them, as described in [23], as a *document-centric* KB which comprise disambiguation-relevant entity information extracted by the original data sets. Formally, we define a document-centric KB as

$$KB_{\text{doc}} = \{d_0, ..., d_n | d_i \in D, n \in \mathbb{N}\} \quad (1)$$

An entry d_i in a document-centric KB consists of the title, the content, both representing a text string, and a set of annotations $\{(t_i, \Omega_i)\}$. An annotation contains a surface form t and a set Ω with entity identifiers. These entity identifiers are referred by the respective surface form t. In the following, we denote an entry in a document-centric KB as

$$d_i = (Title, Content, \{(t_1, \Omega_1) \dots (t_k, \Omega_k)\}) \quad (2)$$

2.2.2 Modeling User Data

In our work the set of user annotations in natural-language documents is called user data [23]. A user annotation consists of a textual representation t, the surface form, and an entity set Ω, which is referred by surface form t. Example 3 shows an annotation of surface form "H1N1", with the *id* denoting an entity's LOD resource:

...<e id="UMLS:C1615607:T005:diso">H1N1</e>... (3)

As depicted in Figure 1 documents with user data from different domains may be collected and stored in a document-centric KB. In our work we assume that user data is readily available and provided by the underlying data set (cf. Section 4).

3. APPROACH

Our document-centric disambiguation algorithm is based on the approach proposed in [23] and can be described as a retrieval based approach for disambiguating arbitrary entities e_i. Given a document-centric knowledge base KB_{doc} containing all entity candidates, a surface form t as well as its surrounding context words c_t^λ (λ denotes the number of words in front of and after surface form t), we return a ranked list R of entities in descending score order, i.e.

$$R = rank(KB_{doc}, t, c_t^\lambda) \tag{4}$$

Our document-centric disambiguation algorithm can be subdivided in two major parts, the search part and the classification part. The approach is similar to a K-Nearest-Neighbor classification using majority voting.

In the **first step** we search for a predefined number τ of documents in our document-centric KB that contain similar content as given by the surface form t and its surrounding context c_t^λ. For this purpose we use a ranking approach, namely Learning To Rank (LTR) [5], to create a ranked list of documents T_τ that contain matching content. More specific, we use a linear combination of a weighted feature set F to compute a score S_{d_i} for each document:

$$S_{d_i} = w^\intercal f_{d_i}(t, c_t^\lambda) \tag{5}$$

Variable w denotes the weight vector for our feature set and function $f_{d_i}(t, c_t^\lambda)$ returns a vector containing the feature values of document d_i with reference to surface form t and its context c_t^λ. Our feature set comprises string similarity features only. We apply the Vector Space Model with TF-IDF weights and the Okapi BM25 model to compute the similarity between the surface form as well as surrounding context and the documents' content [10]. The TF-IDF and BM-25 weights of surface forms and surrounding context are computed with respect to all document titles and contents in the KB. This makes 4 features overall, but our approach leaves the option of choosing other metrics open.

The **second step** encompasses the classification step. We compute the score S_{e_i} for all referenced entities K in our queried document set T_τ:

$$S_{e_i} = \sum_{j}^{T_\tau} p(e_i|d_j) \tag{6}$$

Probability $p(e_i|d_j)$ denotes the probability of entity e_i occurring in document d_j (with reference to all documents in KB_{doc}). To determine the probabilities we apply a modified version of the Partially Labeled Latent Dirichlet Allocation approach (PLDA) proposed in [13], which is similar to the approach of mining evidence for entity disambiguation [8]. By using this approach we are able to consider multiple correct entity references per surface form. However, due to space constraints we omit the LDA model and refer the reader to the referenced paper for details [13]. Again, the result list R consists of the Top-N scored entities. The quality of the results strongly depends on the number of annotated entities in the document set. Generally, user data must be available in our approach.

4. DATA SETS

In the following we present the data sets that are used in our evaluation, namely CALBC[1], Wikipedia and IITB[2].

CALBC: To analyze the user data influence, we use the CALBC (Collaborative Annotation of a Large Biomedical Corpus) data set, a biomedical domain specific KB representing a very large, community-wide shared, silver standard text corpus annotated with biomedical entity references [6]. Overall, we applied the CALBC due to the following reasons:

- In contrast to gold standard corpora like the BioCreative (II) corpora[3], CALBC provides a huge set of annotations which perfectly suitable for our evaluation purpose in terms of quantity (24,447 annotations in Biocreative II versus ≈120M annotations in CALBC). Despite some annotations might be erroneous (silver standard) the corpus most likely serves as a predictive surrogate for a gold standard corpora [6].
- It already represents a document-centric KB comprising documents annotated with biomedical entities.

Basically, the data set is released in 3 differently sized corpora: small, big and pilot. For our evaluation we use the small (CALBCSmall, 174.999 documents) corpus, which differs in the number of available documents. All CALBC documents cover Medline abstracts of the "Immunology" domain, a reasonably broad topic within the biomedical domain. All referenced entities are categorized into four main classes (subdomains) namely, Protein and Genes, Chemicals, Diseases and Disorders as well as Living Beings. These entities are separated in different namespaces. However, the resources of the namespaces UMLS[4], Uniprot[5], Disease (is a subset of UMLS), EntrezGene[6] and Chemlist[7] represent the majority (≈90%) of all entities within the corpus. The UMLS dataset is a combination of many health and biomedical vocabularies, whereas Uniprot provides high-quality resources of protein sequences and function information and EntrezGene exclusively comprises gene-specific information. ChemList is a collection of thousands of chemical substances that are regulated in key markets across the globe. Table 1 depicts the most important statistics of our data set.

Due to a comprehensive taxonomy and classification system a surface form provides 9 entity annotations on average. Some of these entity annotations are linked via same-as relationships in their respective knowledge base.

Table 1: Important Statistics of CALBCSmall

	CALBCSmall
Documents	174.999
Surface Forms	2.548.900
Unique Surface Forms	50.725
Annotated Entities	37.309.221
Unique Entities	453.352

[1]http://www.calbc.eu/
[2]http://www.cse.iitb.ac.in/~soumen/doc/CSAW/
[3]http://www.biocreative.org/news/biocreative-ii/
[4]http://www.nlm.nih.gov/research/umls/
[5]http://www.uniprot.org
[6]http://www.ncbi.nlm.nih.gov/gene
[7]http://www.cas.org/content/regulated-chemicals

Wikipedia: Wikipedia is a well-known, free-access and free content Internet encyclopedia that contains 4.7 million articles in its English language version. By providing interlinks between pages, Wikipedia constitutes a document-centric KB. In contrast to domain-specific KBs like CALBC, Wiki-pedia comprises general knowledge entities but also entities from specialized domains (≈100.000 biomedical entities [17]). In contrast to CALBC, surface forms in Wikipedia refer to one specific identifier instead of multiple identifiers.

IITB: To provide a brief comparison to another document-centric or entity-centric disambiguation approach we use the IITB data set. The IITB data set is the same as used in Han et al. [3]. Overall, it contains 107 web documents. For each document, the surface forms' referent entities in Wikipedia are manually annotated to be as exhaustive as possible. In total, 17,200 name mentions (person names, organizations, objects, abbreviations etc.) are annotated, with 161 name mentions per document on average. In our experiments, we use only the surface forms whose referent entities are contained in Wikipedia [3]. Again, each surface form refers to one specific entity.

5. EXPERIMENTS

Our approaches are implemented in Java with all queries being executed with Apache Lucene 4.8[8]. For the LTR algorithm we chose Sofia-ml[9], a machine learning framework providing algorithms for massive data sets [5]. These algorithm are mainly embedded in our publicly available disambiguation system *DoSeR*[10] (**D**isambiguation **o**f **Se**mantic **R**esources) which is being developed continuously.

First, we briefly explain the parameter settings that are used in our approach (Section 5.1). Second, we investigate the **user data influence**, i.e. how different scales of user data and different values of parameter τ affect the disambiguation results (Section 5.2). In this chapter our intention was not to compare our approach with other approaches because most publicly available biomedical entity annotators do not return a ranked list (e.g. NCBO annotator[11]), which is a key factor in our evaluation. Third, we analyze the **domain dependency**, i.e. how our document-centric disambiguation approach performs on Wikipedia (Section 5.3). The CALBC and Wikipedia data sets serve as documents for our KBs as well as evaluation data sets, while the IITB data set serves for evaluation purpose only. Concerning CALBC and Wikipedia we do not disambiguate and evaluate every single surface form. Instead, we create randomly generated sets of 100.000 surface forms across all documents for each data set. Our document-centric KB comprises either all CALBC documents (Section 5.2) or all Wikipedia documents (Section 5.3), depending on the evaluation.

In our evaluation we report a set of comprehensive measures. This includes the mean reciprocal rank (MRR), recall and mean average precision (MAP) for the CALBC experiments, due to several existing, valid disambiguation results (as provided by the CALBC data set) instead of only one. All these measures are averaged over 5-fold cross validation runs. The Reciprocal rank is the multiplicative inverse of the rank of the first correct result in a result set. Average precision denotes the average of all precision@n values of a single disambiguation task. A precision@n value is computed at every correct hit n in the result set. Unfortunately, measures like discounted cumulative gain (DCG) cannot be applied in our work. These measures are well suited for ranking algorithms, but relevance scores of relevant entities are required, which is not available in CALBC. For data sets which provide only one valid entity per surface form (Wikipedia and IITB), we report the Precision, Recall and F1 values, aggregated across surface forms (micro-averaged). A detailed overview of the used measures can be found in [10].

[8]http://lucene.apache.org/
[9]http://code.google.com/p/sofia-ml/
[10]http://purl.org/eexcess/components/research/doser
[11]http://bioportal.bioontology.org/annotator

5.1 Basic Parameter Settings

Our document-centric disambiguation approach offers several parameter settings, but we only focus on the most important ones. The context length parameter specifies the amount of words of a surface forms' surrounding context in both directions, before and after the surface form. In our evaluation we use a context length of 50 words which provides the best results overall. The context parameter was explicitly evaluated in our previous works [22, 23]. Lower values lead to missing context evidence and higher values lead to additional noise. This observation holds for all experimental parameter and data set combinations (document number and annotation degree).

As a result of using Apache Lucene's TF-IDF score, we note that Lucene's default TF-IDF score also takes internal parameters like term boosting and coordination factor into account. Basically, we always use term queries for surface forms and context terms. The settings for parameter τ are described in the respective sections since this is a parameter we want to analyze. Finally, in our evaluation on the CALBC data set our result list is trimmed to 10 entities per query to provide a good relation between recall and precision. When evaluating against Wikipedia or the IITB data set we return only the entity with the highest score.

5.2 Influence of User Data on Document-Centric Disambiguation

In this experiment we investigate the issue of user data influence. More specific, we analyze how the number of documents used to classify entities (parameter τ) and the amount of annotations within these documents influence the results on CALBC. In the default experiment configuration all annotations in CALBC are used and stored in our document-centric KB (100% user data). For all other scales the KB and probabilities (needed to compute Equation 6) were reconstructed accordingly. To create our KB with a specific fraction of the original user data (for instance 0.1%), we store a user annotation of a CALBC document with the probability $p = \frac{fraction}{100}$ in our KB. The variable $fraction$ denotes the user data fraction with reference to the original data set. Figure 2 and Table 2 show an overview of our results. We note that the plot's x-axis starts at 0.1% due to the necessity of user data in our approach.

Basically, all result values improve with an increase of user data regardless of τ. Poor MRR, Recall and MAP values (between 15% and 35%) when using $\tau = 100$ and 0.1% user data indicate that the amount of user data is absolutely insufficient to provide enough evidence for good disambiguation results. Consequently, using more documents (e.g. $\tau = 1500$) with few annotations per document im-

Table 2: Document-Centric Disambiguation accuracy with various amount of user data

	MRR				Recall				MAP			
UserData in %	100	20	1	0.1	100	20	1	0.1	100	20	1	0.1
$DC_{\tau=100}$	**80.6**	**79.2**	64.1	35.5	**71.7**	**70.3**	51.1	23.1	**59.5**	**58.1**	40.0	16.3
$DC_{\tau=250}$	79.0	78.8	70.8	42.5	68.6	69.0	55.3	28.9	57.8	57.1	45.5	21.6
$DC_{\tau=750}$	76.9	78.2	72.4	53.4	67.3	66.8	58.1	38.8	56.1	55.6	47.7	29.9
$DC_{\tau=1500}$	73.1	75.6	72.4	57.1	65.4	65.2	58.9	42.2	55.3	55.3	48.3	33.7
$DC_{\tau=2500}$	66.4	73.5	**73.9**	**62.4**	60.3	63.4	**59.2**	**46.4**	52.9	54.8	**50.7**	**35.9**

proves the results. The story looks different if too many annotations are available. This is the case when we choose high values for parameter τ and the selected documents contain a high number of annotations on average. While the Recall and MAP values nearly stay constant with $\tau = 1500$ and 100% user data, the MRR significantly drops about 8% due to much noise compared to $\tau = 1500$ and 12% user data. The results attained with $\tau = 2500$ confirm that richly documents in combination with many documents (high values for τ) mitigate the results. However, we dig deeper into this outcome and investigate disambiguation results after determining a fixed number of annotations used for classification (step 2 in the algorithm). Therefore, we introduce a latent parameter Λ, that describes this fixed amount of annotations. Hence, our parameter τ depends on the number of annotations in the documents used for classification and parameter Λ. A first experimental run shows an improvement but does not always provide the best results. This is the case if Λ is set to a low value and the documents used for classification are long and provide many annotations. We also note decreasing results if we use higher values for Λ and use short documents. Again, we have to adjust a parameter according to the underlying data set. We omit an in-depth elaboration of this parameter due to its marginal improvements and space constraints.

In summary, we state that the number of documents used for classification (parameter τ) and the amount of annotations within these documents must be well-matched to attain the best result. An in-depth analysis of the underlying data set is strongly required to attain the best disambiguation results. In future we attempt to provide a quantifiable mathematical formula to yield the optimal parameter values to attain the best disambiguation results based on the data sets' structural properties (e.g. length, annotation frequency).

5.3 Document-Centric Disambiguation on Wikipedia

In the following we expand our evaluation to a data set comprising more general knowledge. More specific, we analyze how our document-centric disambiguation approach performs with Wikipedia entities. For this purpose we first create document-centric KBs comprising Wikipedia articles. In contrast to the abstracts provided by CALBC, Wikipedia articles are longer and vary in their length. As first experiments have shown (cf. Section 5.2), the length of articles plays an important role due to long articles potentially contain more annotations. Thus, we distinguish between using Wikipedia articles as a whole and using paragraphs of Wikipedia articles. When using paragraphs we do not store the whole Wikipedia page in our document-

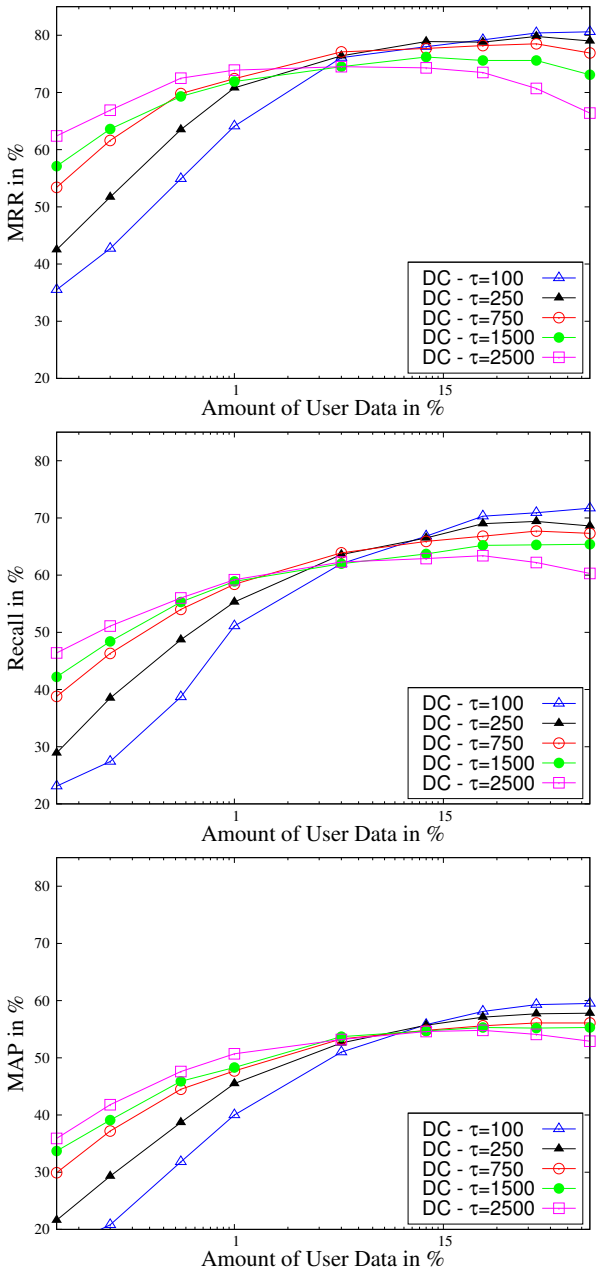

Figure 2: Various scales of user data. Parameter τ denotes the number of documents used for classification.

centric KB but only use one specific paragraph as a KB entry. Consequently, this leads to a significant increase of KB entries (4.3M vs. ≈10M KB entries). It has been shown that search-based document-centric disambiguation is quite robust against large-scale KBs [23] and, thus, we ignore potential result deficits. We apply the Bliki engine parser[12] and extract all paragraphs and their titles, which are are already specified by the Wikipedia authors. In terms of parameter settings we apply $\tau = 100$ in our algorithm because all stored (short and long) Wikipedia documents are richly annotated and this configuration attains the best results with a high number of annotations according to the experiments performed in section 5.2.

In our evaluation we analyze disambiguation results we analyze disambiguation results directly on the Wikipedia and IITB data sets to compare our approach with others. Hence, in the following section we briefly explain existing well-known entity annotators, whose results serve as a good comparison.

5.3.1 Annotators

We evaluate our approach against a strong baseline, two entity-centric and the current state-of-the art collective document-centric disambiguation approach. Unfortunately, all document-centric approaches known to us are not publicly available (e.g. [3, 7, 15]). Additionally, these works often use various and non-consistent data sets, which complicates a comparison. Nevertheless, Han et al. [3] evaluated their approach on the IITB data set and we use their reported results as comparison.

1. **PriorProb**: The Sense Prior $p(e_i|t)$ is generally a strong baseline [2] and estimates the probability of seeing an entity with a given surface form [14]. All probabilities are computed by analyzing available annotated documents. Basically, the PriorProb denotes an entity-centric approach.

2. **DBpediaSpotlight**: Being one of the first semantic approaches (2011) and constituting an entity-centric approach, this framework combines Named Entity Disambiguation and Named Entity Detection based upon DBpedia [11]. Based on a vector-space representation of entities and using the cosine similarity, this approach has a public available web service.

3. **AGDISTIS**: This entity-centric approach [18] of 2014 constitutes an entity-centric, knowledge-base-agnostic approach, based on string similarity measures, an expansion heuristic for labels to cope with co-referencing and the graph-based Hypertext-Induced Topic Search algorithm. The results printed in our work were attained on the DBpedia corpus.

4. **Han et al. [3]**: This document-centric disambiguation approach of 2012 jointly models and exploits context compatibility, the topic coherence and the correlation between them. The algorithm uses words and the mentions in a document as well as global knowledge, i.e. topic knowledge, the entity context knowledge and the entity name knowledge. It is the current state-of-the-art algorithm on the IITB data set.

[12]https://code.google.com/p/gwtwiki/

Table 3: Disambiguation accuracy on Wikipedia

	Precision	**Recall**	**F1**
Wikipedia	35.4	35.4	35.4
Wikipedia$_{Para.}$	**55.1**	**55.1**	**55.1**

Table 4: Disambiguation accuracy on the IITB data set

	Precision	**Recall**	**F1**
Search-based DC	40.9	40.9	40.9
Search-based DC$_{\text{Para}}$	51.5	51.5	51.5
PriorProb	71.3	70.7	71.0
DBpedia Spotlight	76.6	60.5	67.6
AGDISTIS	63.7	30.4	40.2
Han et al. [3]	**81.0**	**80.0**	**80.0**

5.3.2 Results

Table 3 shows the results attained with our document-centric disambiguation approach on our randomly generated Wikipedia data set. Wikipedia$_{Para.}$ represents our approach using Wikipedia paragraphs as documents in our KB. We note that Precision, Recall and F1 values are identical since our approach is incapable of detecting unresolvable surface forms. Instead, we annotate every surface form. Basically, all results are poor on the Wikipedia data sets. The result values significantly increase with the Wikipedia$_{para}$ KB, despite our KB contains more than 10 million entries. This implies that our document-centric approach is not robust against long and richly annotated documents. Although we report comparatively good results with more user data in section 5.2, our results decrease. In this case we have to distinguish between more user data in form of many but short documents and few but very long documents. While the CALBC data set consistently comprises short documents, Wikipedia contains some very long and detailed articles. We assume that these documents introduce additional noise.

Although we cannot compare the results of section 5.2 and this section, the question remains why document-centric disambiguation on the CALBC data set performs significantly better compared to using the Wikipedia$_{para}$ KB (both KBs comprise short documents). We assume that this is referable to Wikipedia's polysemic nature. This leads to that our feature set is not able to return appropriate Wikipedia documents to perform a reliable classification. Wikipedia surface forms and its surrounding context (used as features in our algorithm) might not provide enough evidence to select those relevant documents that contain similar content as provided by the query data.

To compare our algorithm with existing approaches we chose the IITB data set which provides enough surface forms. Table 4 shows the results of our approach in comparison to those approaches described in section 5.3.1. To generate the result values we used the GERBIL - General Entity Annotator Benchmark[13] which offers an easy-to-use platform for the agile comparison of annotators using multiple datasets and uniform measuring approaches [19].

A first mention should be made of the fact that the PriorProb baseline attains an F1 value of 70%, which indicates

[13]http://gerbil.aksw.org/gerbil/

that the surface forms of the IITB data set mainly refer to popular instead of specific entities. While our approach Search-based DC_{Para} (51%) outperforms AGDISTIS (40%), the second entity-centric approach DBpedia Spotlight [11] attains significantly better results in terms of 67.8%. We note that non-collective disambiguation approaches often include prior probabilities [14] how often an entity occurs, so as DBpedia Spotlight. This kind of information is not used at all in AGDISTIS and our document-centric approach, which is partly responsible for the poor values. The F1 value of the probabilistic, generative algorithm by Han et al. [3] shows, that the state-of-the-art currently outperforms all other approaches on this data set (80% F1). However, their approach disambiguates all entities within a document collectively while our approach and DBpediaSpotlight disambiguate each surface form separately. Hence, our approach is able to disambiguate single surface forms with short context (e.g. twitter data). Nevertheless, better disambiguation approaches exist to disambiguate general knowledge entities.

In summary, we state that the results of our search-based, document-centric disambiguation approach on Wikipedia entities is not as reliable as on the biomedical entities. Additionally, shorter documents in the KB are better suited than longer documents due to additional noise.

6. RELATED WORK

Entity disambiguation has been studied extensively in the past 10 years. Several works use intensional entity descriptions (entity-centric KBs) provided by high-quality KBs like DBpedia [4, 11, 16, 18]. One of these works is DBpedia Spotlight [11], a framework for annotating and disambiguating Linked Data Resources and is based on a vector-space model and cosine similarity. Hoffart et al. [4] proposed AIDA for named entity disambiguation tasks. It is based on the YAGO2[14] KB and relies on sophisticated graph algorithms. Another approach is LINDEN, a framework to link named entities in text with a KB unifying Wikipedia and WordNet, by leveraging the rich semantic knowledge embedded in Wikipedia [16]. The approach of Usbeck et al. [18] combines the Hypertext-Induced Topic Search (HITS) algorithm with label expansion strategies as well as string similarity measures and outperforms other approaches using DBpedia.

There are some other disambiguation approaches that use extensional entity-descriptions (document-centric KBs). For instance, Kataria et al. [7] proposed a topic model (Wikipedia-based Pachinko Allocation Model) that uses all words of Wikipedia to learn entity-word associations and the Wikipedia category hierarchy to capture co-occurrence patterns among entities. The authors of [3] also propose a generative approach which jointly models context compatibility, topic coherence and its correlation. Another document-centric disambiguation approach was presented by Sen et al. [15]. Their topic model keeps track of the context of every word in the knowledge base; so that words appearing within the same context as an entity are more likely to be associated with that entity. However, a limitation in their document-centric KB is that each document must describe one specific entity. All these works apply a probabilistic, generative disambiguation model. In terms of search-based document-centric algorithms, the work of Zwicklbauer et al. [22] is the only one found in literature. The authors proposed a non-machine learning entity disambiguation approach, which attains very strong results in the biomedical domain. Zwicklbauer et al. extended their work in [23] and combined document-centric and entity-centric disambiguation to combine the advantages of both worlds.

Biomedical entity disambiguation has also attained much attention in research in the last decade [24]. For instance, Wang et. al classify relations between entities for biomedical entity disambiguation [20]. Biomedical entities can also be disambiguated with the help of species disambiguation. The authors of [21] apply language parsers for species disambiguation and attain promising results.

[14]http://www.mpi-inf.mpg.de/yago-naga/yago/

7. CONCLUSION AND FUTURE WORK

In this work we analyze search-based entity disambiguation with document-centric KBs. More specific, we investigate how the number of documents used for classification and the amount of annotations within these documents correlate and separately influence disambiguation results in the biomedical domain. In this context, we show that both parameters must be well-matched to attain the best result. Further, we expand our evaluation to the more general data set Wikipedia. Our results indicate that search-based entity disambiguation with document-centric KBs performs poorly on Wikipedia. Additionally, the results show that disambiguation accuracy increases when using short documents (e.g. Wikipedia paragraphs) instead of long article pages.

In summary, we state that entity disambiguation with a search-based, document-centric algorithm attain strong results in the biomedical domain. Entity disambiguation on Wikipedia should be performed with entity-centric KBs or generative document-centric approaches which already attain very strong results in other works.

In future work we are going to detect the best para-meter settings for search-based document-centric disambiguation automatically. Additionally, we provide a more detailed evaluation on short data sets like twitter tweets where collective disambiguation is not feasible. On the basis of the results attained in this work we are going to combine entity-centric and document-centric KBs within a federated disambiguation approach to yield the advantages of both. The main goal is to attain strong results on specialized and general domains within one system by choosing the best disambiguation settings automatically.

8. ACKNOWLEDGMENTS

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601.

9. REFERENCES

[1] N. Charbel, J. Tekli, R. Chbeir, and G. Tekli. Resolving XML semantic ambiguity. In *Proceedings of the 18th International Conference on Extending Database Technology, EDBT 2015, Brussels, Belgium, March 23-27, 2015.*, pages 277–288, 2015.

[2] Z. Guo and D. Barbosa. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 499–508. ACM, 2014.

[3] X. Han and L. Sun. An entity-topic model for entity linking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 105–115. Association for Computational Linguistics, 2012.

[4] 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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[5] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142, New York, USA, 2002. ACM.

[6] S. Kafkas, I. Lewin, D. Milward, E. van Mulligen, J. Kors, U. Hahn, and D. Rebholz-Schuhmann. Calbc: Releasing the final corpora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012.

[7] S. S. Kataria, K. S. Kumar, R. R. Rastogi, P. Sen, and S. H. Sengamedu. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1037–1045, New York, USA. ACM.

[8] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1070–1078. ACM, 2013.

[9] F. Mandreoli and R. Martoglia. Knowledge-based sense disambiguation (almost) for all structures. *Information Systems*, 36(2):406–430, 2011.

[10] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[11] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, USA. ACM.

[12] C. Ogden and I. A. Richards. The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism. *8th ed. 1923. Reprint New York: Harcourt Brace Jovanovich*, 1923.

[13] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[14] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[15] P. Sen. Collective context-aware topic models for entity disambiguation. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 729–738, New York, NY, USA, 2012. ACM.

[16] W. Shen, J. Wang, P. Luo, and M. Wang. Linden: linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 449–458, New York, NY, USA, 2012. ACM.

[17] L. Tian, W. Zhang, A. Bikakis, H. Wang, Y. Yu, Y. Ni, and F. Cao. Medetect: A lod-based system for collective entity annotation in biomedicine. In *Web Intelligence*, pages 233–240. IEEE Computer Society, 2013.

[18] R. Usbeck, A.-C. Ngonga Ngomo, S. Auer, D. Gerber, and A. Both. Agdistis - graph-based disambiguation of named entities using linked data. In *Int. Semantic Web Conf.* 2014.

[19] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – general entity annotation benchmark framework. In *24th WWW conference*, 2015.

[20] X. Wang, J. Tsujii, and S. Ananiadou. Classifying relations for biomedical named entity disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1513–1522, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[21] X. Wang, J. Tsujii, and S. Ananiadou. Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, 26(5):661–667, 2010.

[22] S. Zwicklbauer, C. Seifert, and M. Granitzer. Do we need entity-centric knowledge bases for entity disambiguation? In *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies*, i-Know '13, pages 4:1–4:8, New York, NY, USA, 2013. ACM.

[23] S. Zwicklbauer, C. Seifert, and M. Granitzer. From general to specialized domain: Analyzing three crucial problems of biomedical entity disambiguation. In *International Conference on Database and Expert Systems Applications - DEXA 2015*, 2015.

[24] S. Zwicklbauer, C. Seifert, and M. Granitzer. Linking biomedical data to the cloud. In *Smart Health*, pages 209–235. Springer, 2015.